CRANFIELD UNIVERSITY

TIMOTHY S. FAREWELL

INNOVATIVE METHODS FOR SOIL PARENT MATERIAL MAPPING

SCHOOL OF APPLIED SCIENCES

PhD THESIS

ACADEMIC YEAR: 2009-2010

SUPERVISOR: T. MAYR

JANUARY 2010

TIMOTHY S. FAREWELL

INNOVATIVE METHODS FOR SOIL PARENT MATERIAL MAPPING

SUPERVISOR: T. MAYR

JANUARY 2010

ABSTRACT

Soil parent material exerts a fundamental control on many environmental processes. Nevertheless, resulting from the separate mapping programmes of the geological and soil surveys, parent material is currently poorly mapped in the United Kingdom. This research develops and tests four methods of predicting soil parent material using three study areas in England. The qualities of desirable parent material maps were stated, and then used to create new map value metrics to assess the success of the four methodologies.

Firstly, translations of surface and bedrock geology maps to parent material maps were tested, using international and national parent material classifications. Secondly, qualitative expert knowledge of parent material, captured from published literature, was formalised into inputs for a corrected probability model. Parent material likelihood was predicted using three map evidence layers: geology, slope and soil. Thirdly, extensive data mining was used to create fully quantitative inputs for the same probability model, and the results were compared. The final method provided a quantitative framework for the expert knowledge model inputs by the incorporation of sparse data sampling.

The expert knowledge method created parent material maps of higher value than those created by the translation of geological maps. However, the inputs derived from qualitative expert knowledge were demonstrated to benefit from the addition of quantitative sample data. The resulting maps achieved overall accuracies between 60% and 90% and contained numerous detailed classes with explicit probabilities of prediction. Extensive parent materials were shown to be predicted well, and physically and chemically distinctive parent materials could be effectively predicted irrespective of their extent. Parent material class confusion arose between units where the evidence datasets were unable to provide the sufficient geographic or descriptive detail necessary for differentiation. In such cases, class amalgamation was used to overcome consistent misclassification. Recommendations are provided for the application of this research.

EXECUTIVE SUMMARY

This executive summary seeks to provide the reader with an overview of the numerous methodologies and approaches used in this research. It can be used to provide quick reference for how each method or section fits within the research.

Detailed and high quality maps of soil parent material are required as a data input to models to help address a growing range of environmental issues in England and Wales.

As soil parent material occupies a position between soil and drift or bedrock geology, parent material maps have often been created from existing soil or geological mapping. Some parent material maps have been created using remote sensing, but many of these approaches are not applicable in England and Wales due to the temperate climate and thick vegetation cover. These limit bare, dry soil conditions which are ideal for remote sensing approaches. A few remote sensing techniques such as gamma radiometrics can be interpreted to provide promising predictions of parent material. However, they are expensive to run and, so far, the datasets have limited availability in the UK.

Creating soil parent material maps from geology maps is not straightforward as geological maps tend to be chronostratigraphic. Therefore clear statements of the lithology of the units may not be readily available. Furthermore the extent and quality of the mapping of superficial deposits can be inconsistent. Thus geological maps can be of limited value for predicting classes of soil parent material.

Approximately 30% of England and Wales is covered by detailed soil maps. From these detailed sources of soil information, robust soil parent material maps can be derived using defined translations between soil classes and parent material types. Such areas are suitable for the testing and development of methods of creating parent material maps in currently unmapped areas.

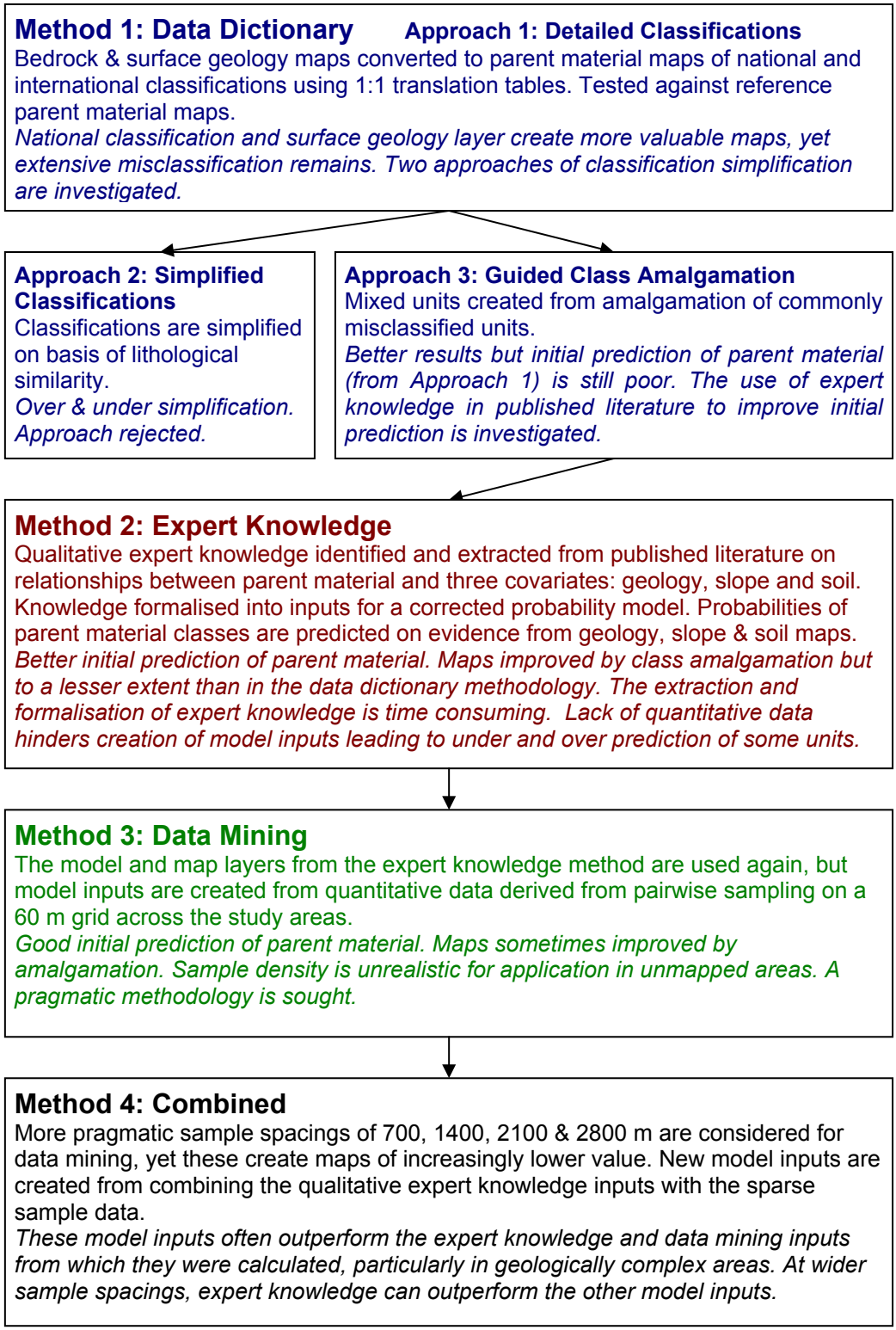


Figure 1 – Roadmap of methods used in this research

Three study areas were chosen to test the success of four methods of modelling soil parent material. The areas represent some of the different soil and geological characteristics found in England. The study areas have been mapped in detail with soil maps published at 1:25,000 scale. Reference soil parent material maps were made for each of these study areas by interpreting and translating from the soil map legend.

An approach to creating parent material maps is the reinterpretation of existing geological mapping. To test how successful this might be, a data dictionary approach was used to convert existing geological maps to parent material maps using one-to-one (geology to parent material) translation tables. Both national and international parent material classifications were used to create parent material maps, using both bedrock geology and surface geology inputs. The classifications and inputs were tested to see which produced the highest quality maps.

Assessing the quality of the resulting parent material maps proved difficult. This was because traditional metrics of map classification success, such as overall accuracy (θ_1) or the kappa statistic (κ), were found to provide only a limited assessment of the value of a map, as broad class definitions could create accurate maps but with very few classes or little class detail. Therefore the desirable attributes of parent material maps were explicitly stated. These include numerous, specific parent material classes related to both geology and soil. These classes would accurately represent the geographic reality, and provide the map with a high overall agreement between the model and reference maps. From these statements, new metrics of class value (ξ) and map value (ψ_3) were derived and have been used to compare the success of models and methods throughout this research.

Initially, due to significant misclassification of the modelled parent material map, neither national nor international classification using either geological input produced satisfactory results. Therefore, two approaches of classification simplification were considered. The first simplified the entire classification on the basis of lithological similarity. While this simplification approach resolved some misclassifications, it oversimplified other parent material units unnecessarily. It therefore did not achieve the

desired levels of improvement over the initial translations. The second approach to classification simplification examined the consistent misclassifications with the reference map, and used these analyses to guide specific class amalgamations on a study area basis. For example, a pebbly sandstone was consistently confused with a pebbly drift, so these units were amalgamated. This approach achieved considerably higher map values (ψ_3) than either the initial translation or the simplification of the entire classification.

It was shown that the national parent material classification consistently produced parent material maps with higher value than those using the international classification. Furthermore it was demonstrated that surface geology maps substantially out-performed bedrock maps in areas with extensive superficial deposits.

Despite the improvements in map value produced by class amalgamation, it was felt that attempts to improve the prediction of parent material, prior to amalgamation, were warranted to attempt to further improve map detail. A second methodology attempted this using available expert knowledge combined with additional environmental datasets.

A thorough review was undertaken of potential extra sources of information that could be incorporated in order to improve the prediction of soil parent material. Qualitative expert knowledge on the relationships between soil parent material and environmental parameters such as geology, soil and slope, is held within published literature and national soil and geological databases. This qualitative knowledge was extracted, structured, assessed and formalised to create pseudo-quantitative inputs for a corrected probability model.

Much of the identified expert knowledge was inconsistent in the breadth and depth of the descriptions of the environmental relationships. Nevertheless, enough knowledge was contained within published literature to build models with three evidence layers. These described the relationship between parent material and the slope class (SLOPE), surface geology (GEOLOGY) and regional soil associations (SOIL).

Models were run with all combinations of these three inputs. It was shown that the use of expert knowledge could result in a dramatic increase in the value of the resulting parent material maps over those produced by a one-to-one translation, both prior to and following class amalgamation. The data dictionary (first) method predicted one parent material class for each geological unit. The expert knowledge (second) method provided the probability of each parent material class, for each evidence layer combination.

Despite these improvements, concerns remained that qualitative descriptions were being used to create quantitative inputs for a probability model. Therefore a third method employed pairwise data mining techniques to test the efficacy of a quantitative approach. The same probability models were populated with inputs derived from pairwise analysis of the relationships between parent material classes and the classes of the predicting evidence layers.

This model with fully quantitative inputs was trained and tested using a dense 60 m grid over the entirety of each study area, providing a high map value (ψ_3) for comparison with the other methodologies. The GEOLOGY and SOIL inputs were shown to be better predictors of parent material than SLOPE. The data mining method produced the most valuable maps, but the input data was trained from an unrealistically detailed sampling of the study area (300 points per km^2) and tested on the same area. Because of concerns about the over-optimisation of models trained and tested on the same area and the unrealistic sample density, a fourth and final methodology combined aspects from the expert knowledge and data mining methodologies to test more pragmatic applications of this research.

Firstly, a range of data mining models were run using increasingly sparse data samples, collected on 700, 1400, 2100 and 2800 m grids. Secondly, these same samples were used to provide a quantitative framework for the qualitative expert knowledge. Conditional probability tables for each ‘evidence layer: parent material’ pair were derived from the expert knowledge and the sparse data sample. Aspects of each were used to create new quantified expert knowledge conditional probability tables. These were used as inputs into further models.

The value of the maps created with the increasingly sparse data samples decreased as sample spacing increased. The inputs combining expert knowledge and the data sample had a similar trend, but tended to produce maps of higher value than those achieved by just the quantitative data, particularly in the more complex areas.

Extensive parent material units were typically better predicted than units with limited extent. However, distinctive parent material units such as chalk, peat and alluvium, were shown to be better predicted than would be expected by their limited extent. Such units are easily recognisable by both geological and soil surveyors.

Parent materials relating to thick drift deposits and bedrock geology were consistently well predicted, but it was shown that the evidence layers (GEOLOGY, SOIL, SLOPE) struggled to accurately predict certain parent materials. These included those parent materials derived from thin drift, or where differentiating characteristics such as subtle changes in the stoniness occur in the top 45 cm of the soil profile. Such predictive inaccuracies typically result from different mapping priorities of the geological and soil survey, and the different mapping scales used on evidence and reference maps. The addition of further detail to the classification and linework represent areas for future research.

Overall map accuracies (θ_1) based on the most likely predicted parent material ranged between 60% and 90%. The higher level represents a very useful parent material input for a number of environmental and soil models. Furthermore, it has been shown that where the most likely parent material did not agree with the reference parent material map, it was common that the second or third most probable parent material was in agreement. In such cases, the use of amalgamated classes to deal with class confusion can be recommended. In all cases the probability of each parent material class can be used as an assessment of the confidence of the mapping, allowing propagation of the knowledge of errors into future work. As such these innovative approaches offer a promising method for the creation of useful parent material maps for England and Wales.

ACKNOWLEDGEMENTS

Many people have supported me during this PhD, and I should like to acknowledge their assistance.

In particular, I should like to thank my supervisor, Thomas Mayr for his helpful advice and guidance throughout the six year course of this research. Many thanks also to Tim Brewer and Bob Palmer for their insight and support.

Thanks also go to my colleagues at NSRI, both past and present, who have provided practical and moral support. Thanks to Andrew Rayner, for your assistance with the VBA coding, and to Bob Jones for our many conversations. Thanks also to Ian Truckell for your practical help and to Caroline Keay and Sara Larman who have been my very supportive office mates throughout this PhD.

To my brother, Daniel Farewell, thank you for your general and specific assistance.

To my parents, thank you for all your consistent love and advice.

To my Lara, thank you for your love and sacrificial support throughout the entirety of this project. Ana and Rose, thank you for the sheer joy you bring to me through just being yourselves.

CONTENTS

FIGURES

TABLES

SYMBOLS AND ABBREVIATIONS

y	The overall accuracy of the map: the proportion of the map correctly predicted
ξ	Class value metric: ξ is calculated as the geometric mean of the user (A_u) and producer (A_p) accuracies for the parent material unit in question.
ψ_3	The map value metric - This metric considers the specificity, accuracy and number of predicted map units as well as the overall accuracy.
ω	The weighted class value metric
C_t	Total number of parent material classes identified in either reference map or modelled map
Ce	The number of parent material classes in both reference and modelled maps
AEM	Airborne electromagetics
BGS	British Geological Survey
Class Value	(ξ) is calculated as the geometric mean of the user (A_u) and producer (A_p) accuracies for the parent material unit in question.
DEM	Digital Elevation Model
DSM	Digital Surface Model
DTM	Digital Terrain Model
Effective classes	the number of parent material classes in both reference and modelled maps
EK SLOPE	Slope model input derived from qualitative expert knowledge
EM	Electromagnetic
EMI	Elecromagnetic Induction
ESB	European Soil Bureau parent material classification (used in the data dictionary methodology)
ESB12	The European Soil Bureau parent material classification with dominant (1) and secondary classes (2). (used in the data dictionary methodology)
Evidence layers	GEOLOGY, SLOPE, SOIL
Expector	A software package from which the probability model used in this research was derived.
GEOLOGY	Geological evidence layer, derived from BGS DiGMap50 (digital geological map).
GIS	Geographic Information System
GPR	Ground Penetrating Radar

Kappa	Fleiss's variant of Cohan's kappa statistic: the amount of agreement between the modelled map and the 'truth', minus the chance agreement.
LandIS	The Land Information System for England and Wales
Map value	This metric considers the specificity, accuracy and number of predicted map units as well as the overall accuracy. (see ψ3)
MS	Magnetic susceptibility
NSI	National Soil Inventory (a 5 km grid sampling of the soils of England and Wales)
NSI SLOPE	Slope model input derived from the NSI national survey
NSRI	National Soil Resources Institute of Cranfield University, UK
O.D.	Ordnance Datum (meters above sea level)
Overall accuracy	The proportion of the map correctly predicted
PARENT	A component of the NSRI parent material classification which describes the broad physical nature of the substrate
PARLITH	The full NSRI parent material classification including PM_LITH and PARENT (used in all methods but the data dictionary methodology)
PM_LITH	The lithological component of the NSRI parent material classification (used in the data dictionary methodology)
SLOPE	A slope class layer, derived from NextMap 5 m DTM
SOIL	The National Soil Map
VLF	Very Low Frequency

GLOSSARY OF TERMS

Bedrock	Consolidated, pre-Quaternary Geology
Correlative	In this research, this describes possible surrogate maps for parent material.
Covariate	Describes the evidence layers used to predict parent material (SLOPE, GEOLOGY, SOIL)
DTM derivative	A range of datasets describing the landform may be derived from a digital terrain model. These are DTM derivatives, and include datasets describing slope, aspect and slope curvature amongst others.
Evidence layer	These are the predicting layers: slope, geology map and national soil map
Geo-diversity	The variety of earth materials and processes which are found in, shape and affect a particular area.
Horizons	Layers within the soil profile
Hypothesis	In this research, this refers to a particular parent material class which is being predicted. i.e. *"the hypothesis is that BhB1 is present at this location"*. This is tested in the probability model and compared with all other parent material classes (or hypotheses).
Model input	These refer to the numeric probability tables defined for each parent material / evidence layer pair.
Parent material	The mineral or organic matter from which soil forms, found at the base of the soil profile.
Reference map	Reference parent material maps were derived from detailed soil series mapping of the study areas. These were translated by defined translations to parent material maps
Regolith	A generic term for the loose material which overlies consolidated geological deposits. This includes weathered bedrock material as well as the soil.

Soil profile	The vertical succession of layers or horizons in the soil.
Soil series	A taxonomic unit of soil, differentiated on the basis of the nature of the parent material, textural characteristics, and the presence or absence of material with a distinctive mineralogy or colour.
Superficial deposits	Less consolidated geological deposits typically deposited by glaciers, water or wind. As these deposits are usually terrestrial, they tend to be less continuous geographically than bedrock deposits, and more variable in particle / clast size.
Texture	A description of the particle size distribution of soil and the relative composition of sand, silt and clay particles.

1 INTRODUCTION AND RESEARCH CONTEXT

This chapter sets out the context of this research. It describes the need for soil parent material maps to address a range of environmental issues, from the creation of a continuous near-surface hydrological model to the prediction of detailed soil classes. It discusses some of the differences between maps of soil parent material and those of geology, and suggests that the relatively low use of parent material inputs in soil models arises from these fundamental differences.

1.1 The requirement for soil parent material maps

The soil – geology continuum is complex and heterogeneous, and while maps for both soil and geology exist, in the United Kingdom they were mapped by separate organisations with different mapping priorities. The soil survey investigated the soil to a maximum depth of 1.2 m, while the geological survey tended to not map units which were thinner than 2 or 3 m thick. As a result, the interface between soil and geology has not been effectively mapped. This relatively unmapped interface is the realm of the soil parent material.

The term 'soil parent material' has been used in a number of slightly different ways by different organisations and countries. In England and Wales, and in the context of this research, soil parent material is defined as the mineral or organic matter from which the soil formed, found at the base of the soil profile. Parent materials are differentiated on the basis of the nature of the substrate. For example, all bedrock and skeletal material substrates are termed *lithoskeletal*, while soils with abundant stones are defined as *gravely* or *over gravel*. *Thick drift* substrates and soils with a *soft, pre-Quaternary* substrate are also identified. Apart from the soft, pre-Quaternary substrate type, no distinction is made between parent material units with similar lithologies but different stratigraphical ages. Because of this interaction between aspects of the soil, and

underlying geology, parent material classes are related to, but taxonomically differentiated from the regolith, solid bedrock geology, quaternary, superficial or aeolian deposits and the soil itself. These terms are described in more detail in the Glossary of Terms (p xxiii) and the relationships between these components of the near surface continuum are described in Figure 2 .

While a variety of parent material types exist (Appendix 8), two generic types are presented for comparison in Figure 2. Figure 2 (a) shows a soil developed from a parent material strongly related to the underlying bedrock geology, while (b) shows a superficial deposit overlying solid bedrock. In the case of (a), the bedrock may be considered the parent material of the soil, while in the case of (b), the parent material is the superficial deposit. In some situations, the parent material may be entirely incorporated into the soil, leaving no original parent material between the soil and unrelated bedrock.

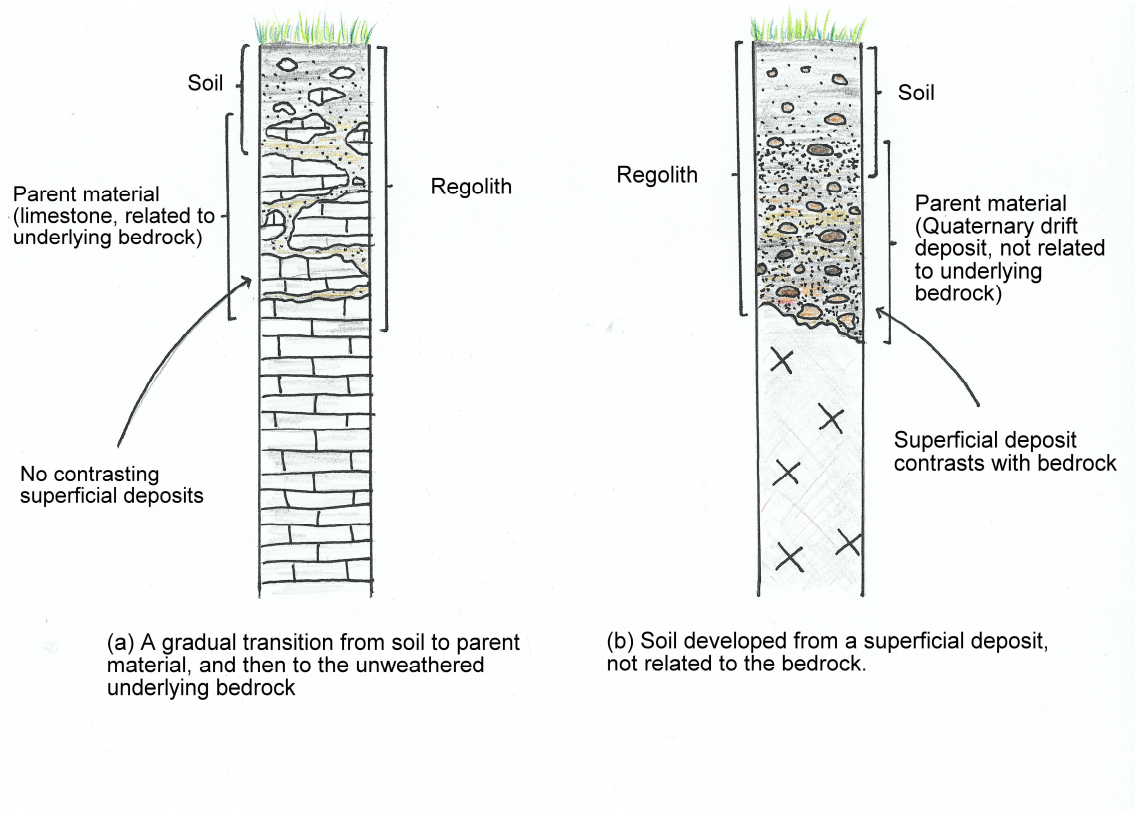


Figure 2 - Schematic diagrams showing the relationship between soil, parent material, superficial and bedrock geology and related terminology.

Parent material is the primary differentiator of taxonomic soil series in England and Wales (Clayden and Hollis, 1984) and is a fundamental component of the national soil classification system. As such, highly detailed soil series maps can be reinterpreted to describe the parent material. However, detailed soil maps (between 1:25,000 and 1:50,000 scale) are available for less than a third of these countries (Figure 3) and the only continuous soil mapping is the 1:250,000 scale National Soil Map (NSRI, 2008a). This map employs mapping units which contain numerous soil series and numerous parent material types, making it unsuitable for the production of high resolution maps of parent material. Detailed parent material maps are required for the two-thirds of England and Wales currently lacking detailed soil maps.

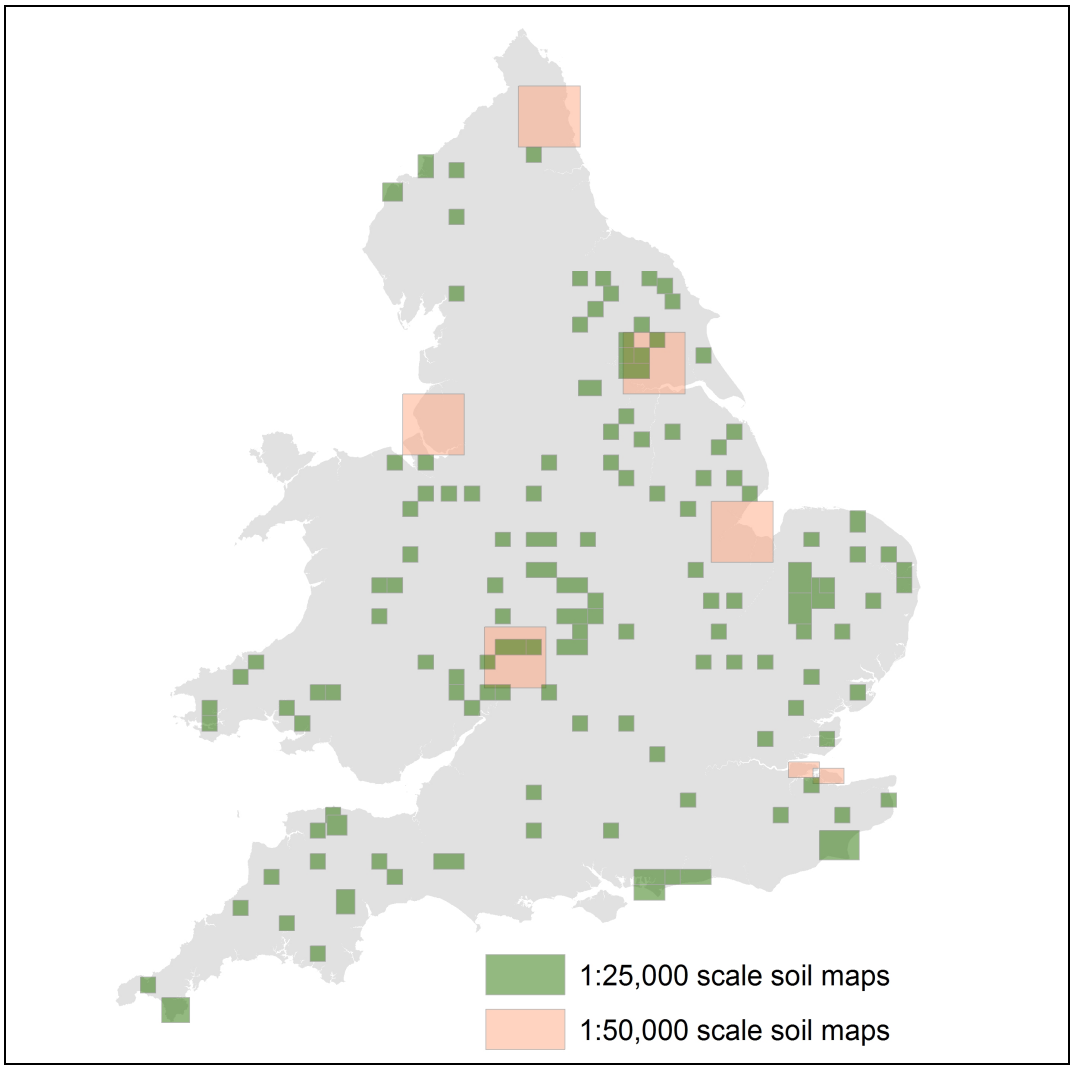


Figure 3 - The availability of detailed soil mapping in England and Wales

Geological mapping is available for all of England and Wales at 1:50,000 scale. However, the physical nature of the soil substrate, including descriptions of gravely layers or thin superficial deposits, which are necessary for accurate descriptions of soil parent material, are often excluded from such mapping. As such, initial attempts to reclassify geological maps to parent material maps have led to extensive misclassification of parent material units (Palmer et al., 2007) and geological maps have been shown to also be imperfect surrogates of soil parent material.

Some requirements for detailed maps of soil parent material in the UK

- An input to predictive models for detailed soil mapping
- Guide to mineral and aggregate extraction
- Enabling the creation of a continuous, near surface hydrological model
- Flood and near surface flow modelling
- Protection of water quality
- Prediction of water chemistry
- Modelling distribution and sources of heavy metals
- Aid in the definition of ecosystem types
- Development and management of habitats
- Improved mapping of geohazards, such as subsidence

While high quality maps describing the nature of the parent materials in England and Wales are currently unavailable, there is a growing need for such maps, particularly at a local scale. Parent material maps are required in their own right to guide mineral and aggregate extraction, and are a vital layer to facilitate the creation of a continuous hydrological model of the Earth's surface (Wysocki et al., 2005). Such models can aid the protection of water quality and prediction of water chemistry (Billett et al., 1997; Grieve, 1999).

Parent material maps can enable enhanced flood and subsurface flow modelling (Mosley, 1982; Bishop et al., 1990), provide information on the distribution and sources of heavy metals (Lado et al., 2008; Manta et al., 2002) or aid in the definition of ecosystem types (Moncoulon et al., 2004). The mapping and improvement of soil fertility and ecosystem services would be aided by good quality maps of soil parent material (Lorenz and Lal, 2009), and such maps would be key to enhanced understanding and application of soil functions. These are wide ranging, including food and biomass production, environmental pathways for water and the location and concentration of pollutants (Blum, 1993; Rodríguez Martín et al., 2006), issues surrounding the development and maintenance of the biological habitat and gene pool (Moles and Moles, 2002), supporting the source of raw materials, such as timber (Frey et al., 2009), protecting and supporting the physical and cultural heritage (Homburg, 2005), and providing a platform for human development (Igué et al., 2004).

Because of the strong chemical link between soil and parent material, it has been proposed that parent material maps be used to guide plant sampling programmes to identify selenium excesses or deficiencies in Canadian prairie crops (Doyle and Fletcher, 1977). In more volatile areas, an understanding of parent material has aided in landmine clearance operations (Hannam and Dearing, 2008) and has guided assessments of the vulnerability of structures to earthquake damage (Northey, 1974). A reliable map of parent material would also improve the existing geohazard (BGS, 2010) and natural perils datasets (NSRI, 2009) currently used by the financial services industry.

Existing soil maps have been converted into parent material maps (Roy et al., 1997; RI USDA NRCS, 2009; European Soil Bureau, 2001). However, as less than a third of England and Wales is covered in detailed soil mapping, this approach is not suitable for much of the area. Additionally, as one of the requirements for parent material maps in the UK is to assist in the production of new detailed soil maps, it is necessary to have a robust methodology for the generation of parent material maps derived from information other than existing detailed soil maps.

Parent material maps can be an important input into predictive environmental models, to predict soil classes and properties as parent material supplies essential information about physical properties of the substrate. Such predictive models are widely used to create soil maps from environmental covariates (McBratney et al., 2003) for unmapped areas or those covered only by reconnaissance scale maps where greater knowledge of local soil distribution is required.

Digital soil mapping offers an alternative and more quantitative method of creating soil maps than traditional soil survey (Rossiter, 2005; Mayr et al., 2001) by using statistical methods in combination with soil observations and a range of environmental covariates (McBratney et al., 2003; Dobos et al., 2006). These models tend to be based around certain components of Jenny's (1941) seminal mechanistic equation of soil formation; that soils are a function of climate, organisms, relief, parent material and time. For the purposes of modelling soils, the location and properties of nearby soils can also be used (McBratney et al., 2003). Compared to relief or digital terrain model derivatives, geological data is rarely used in these models (Figure 4), and pure parent material datasets are even less common.

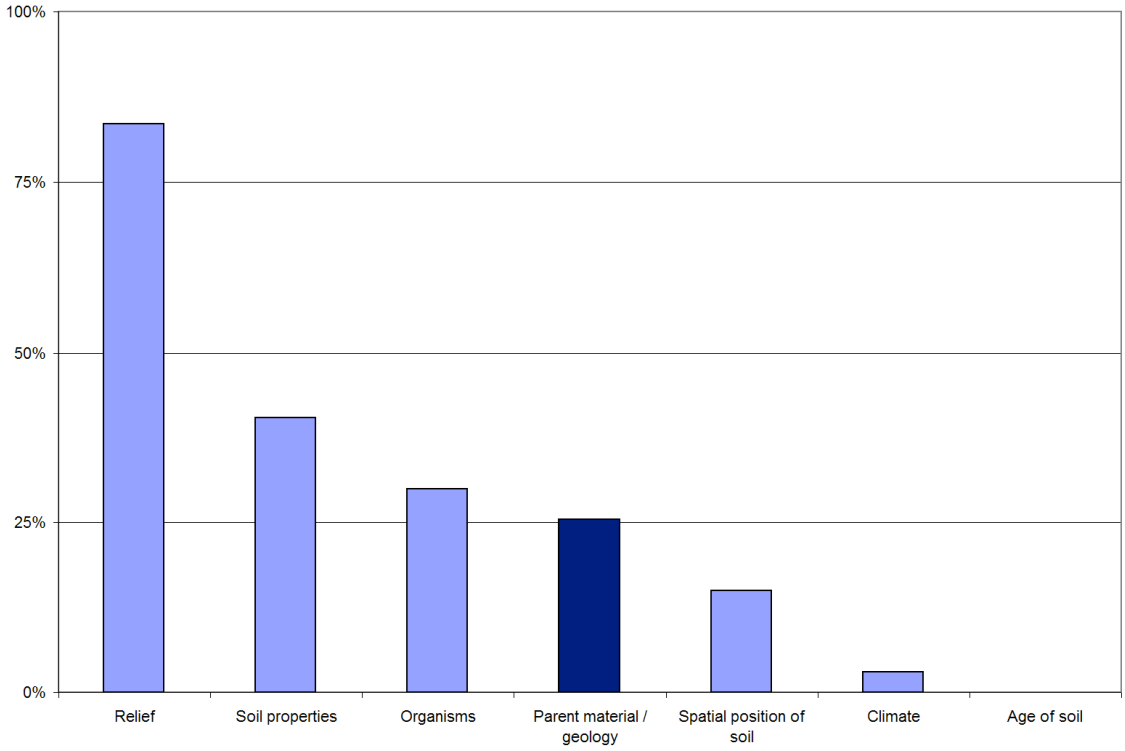


Figure 4 - the percentage of published studies using the seven inputs into digital soil mapping exercises, based on a survey of 132 papers (derived from McBratney et al., 2003)

Studies which have used geological data for digital soil mapping (Thomas et al., 1999a; Thomas et al., 1999b; Bui and Moran, 2001; Bui and Moran, 2003; Ramli, 1996; Cook et al., 1996) are less common than those using relief and landform data, with less than 30% of the studies using parent material inputs (McBratney et al., 2003). Because soil parent material, soil texture and drainage are so intimately related, it is remarkable that so few digital soil mapping programs use a parent material correlative as a model input. This appears to be primarily because geological maps are imperfect surrogates for soil parent material maps.

There are fundamental differences in the mapping priorities, and resulting maps of geological and soil surveys. Soils are described with strong reference to the lithology from which they have formed, whereas geological map units tend to be chronostratigraphic, and can group multiple lithologies of similar ages. Historically, the British Geological Survey has underemphasised the spatial extent of superficial deposits (Palmer et al., 2007) and rarely describes these in detail (British Geological Survey,

2009). This leads to extensive misclassification of certain parent material units (Palmer et al., 2007). These issues make the creation of a parent material map from a geological map challenging.

There are problems arising with the use of traditional geological data as a parent material map. However, due to the strong influence of geology on the texture of the soil, the close relationship between parent material and the taxonomic soil series in England and Wales (Clayden and Hollis, 1984), the national coverage and the ready availability of digital geological data, steps should be taken to transform geology into a parent material correlative.

It has been shown that parent material maps could contribute to a range of hydrological, ecological, economic and sociological applications, yet these maps are not widely available in the UK. Methodologies are therefore required for the creation of detailed parent material maps.

1.2 HYPOTHESIS, AIMS AND OBJECTIVES

1.2.1 Hypothesis

It is hypothesised that, with appropriate techniques, effective maps of soil parent material may be derived from existing sources of geological, soil and landscape information.

1.2.2 Aims

To develop, investigate and evaluate methodologies suitable for the creation of useful soil parent material maps.

To make recommendations as to the effective application and implementation of the results of this research.

1.2.3 Objectives

1. To define the qualities and attributes of soil parent material maps which are useful for addressing a range of environmental applications.

2. To test the use and value of national and international parent material classifications in the correlation of geological and existing soil parent material data.

3. To test the use and value of bedrock geology and surface geology as predictors of soil parent material.

4. To investigate methods of classification simplification for situations where parent material units are often misclassified.

5. To identify and extract the qualitative expert knowledge of soil surveyors contained within published literature, and investigate the use to which this can be put in constructing predictive soil parent material models.

6. To test the use of pairwise data mining procedures to extract and quantitatively assess spatial co-incidence patterns gleaned from existing geological, soil and slope datasets for the purpose of modelling soil parent material.

7. To combine aspects of both quantitative data mining and qualitative expert knowledge to create a quantified expert knowledge soil parent material model.

8. To evaluate the fitness for purpose of derived parent material maps.

9. To make recommendations as to the effective creation of parent material maps for use in environmental modelling.

The focus of this study is not to explore the spatial application, such as extrapolation, of the techniques examined here, but rather, this study explores how to best create a parent material map from readily available datasets.

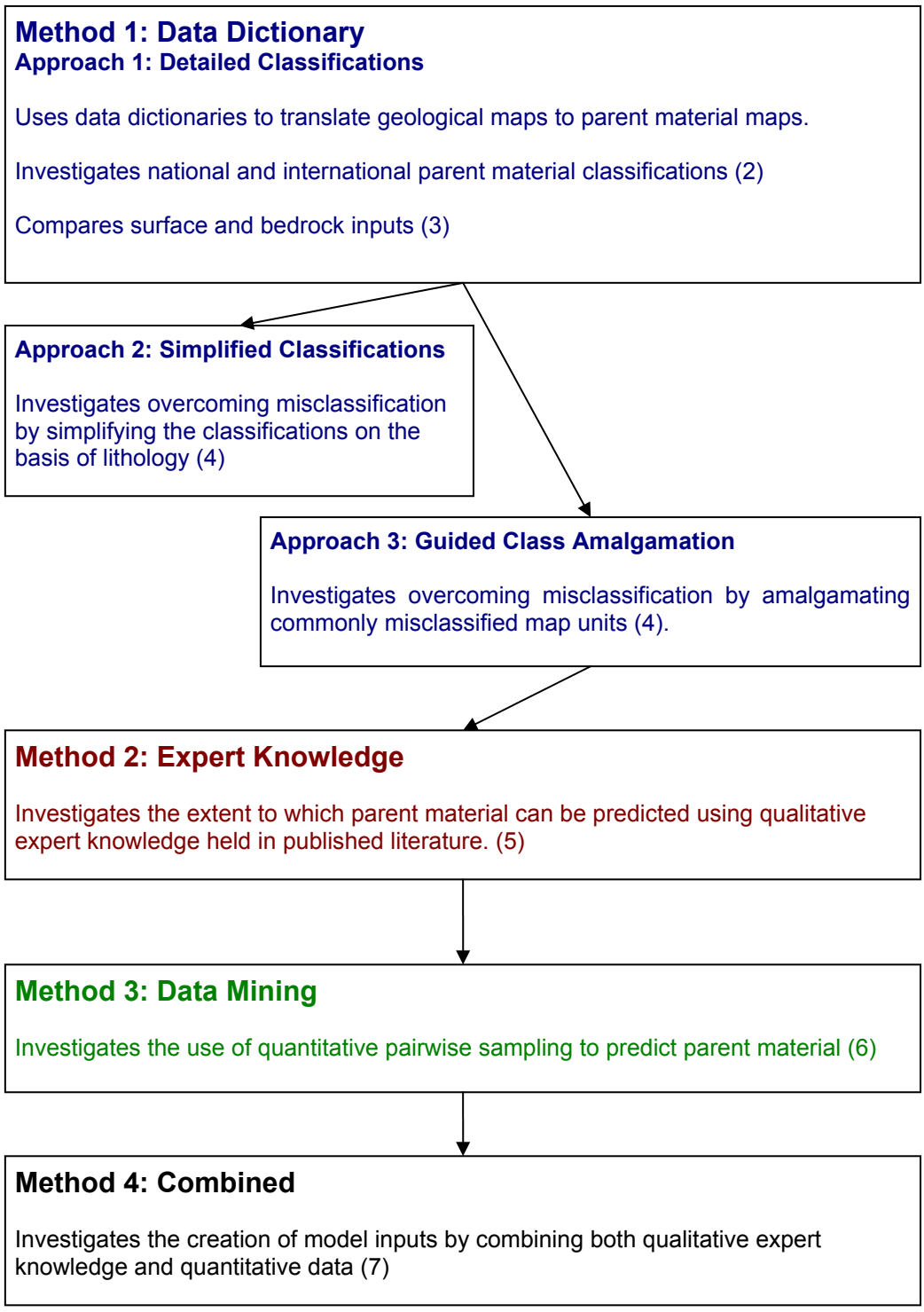


Figure 5 – Simplified roadmap of methods used in this research and the objectives each address.

Note: Objectives are indicated by numbers in brackets. Objectives 1, 8 and 9 are addressed outside these methodologies.

2 REVIEW OF PARENT MATERIAL MAPPING

The mapping or modelling of soil parent material has rarely been undertaken as an exercise in its own right, and there is very little published literature on this subject. Nevertheless, this chapter will discuss published attempts to model or map parent material and will also explore potential surrogate techniques and data inputs which may assist parent material mapping in England and Wales.

2.1 Creating parent material maps from soil survey maps and field data

Soil parent material is the geological material from which soil forms, found at the base of the soil profile. It can be considered as the basic material for pedogenesis. In England and Wales, soil parent material is a fundamental component of the taxonomic soil series definition. Therefore, by using this defined relationship between soil series and parent material, existing soil series maps can be translated to parent material maps. The majority of existing soil parent material maps have been derived from existing soil survey maps. Nevertheless, such translations from soil series to parent material are not possible in regions without detailed soil series mapping. For example, the map units of the 1:250,000 scale National Soil Map of England and Wales (NSRI, 2008a) are soil associations (groupings of soil series). Within these mapping units, reliable information on the presence of particular soil series is unavailable and multiple parent materials are typical. In such cases, the derivation of a parent material map is more complex.

Wysocki et al (2005) recognised the need for an integrated knowledge of the soil-geology continuum to address a number of environmental concerns in the United States, from water quality and nutrient management to landfill placement. While in the States, soil taxonomy is officially constrained at a depth of 2 m, soil surveyors frequently do

record information in notebooks on the deeper layers when describing profile pits or road cuts. Wysocki et al (2005) suggest the use of notes from soil surveys to provide information on the subsurface, though note that such tasks are 'tedious and subject to error'. At a regional scale in Iran, soil profile data with subsurface information was downscaled across an extensive study area to make a very general soil parent material map as a component of a process of understanding pedodiversity (Toomanian et al., 2006).

When field surveys are undertaken, surveyors commonly annotate field maps with information on observations and include diagrams and field sketches on the margins of the map itself and occasionally on the back of the map. Field sheets can therefore contain a wealth of information, but this information requires a significant amount of work to collate, due to its unpredictable and inconsistent nature. Such notes are also often lacking a map key. Additionally, being unique, field sheets are generally very difficult to access.

Wysocki et al (2005) encourage current soil surveyors to make better use of block diagrams and to produce lithostratigraphic, pedostratigraphic, geomorphic and soil parent material maps during the field mapping programme. This approach of specifically noting the parent material whilst in the field has been employed during the Swedish survey of Forest Soils (Odell and Lofgren, 2006) where both soil texture and parent material are assessed in the same 23,500 trial pits. This has resulted in a range of national scale parent material maps describing both the genesis and grain size distribution (Lundin, 2006). These parent material maps are based on samples at 80 cm depth (Olsson, 1999) and describe the texture of the parent material and one of five major classes reflecting the mode of deposition, namely: sediments with high degree of sorting; sediments with low degree of sorting; till; bedrock outcrop; and peat.

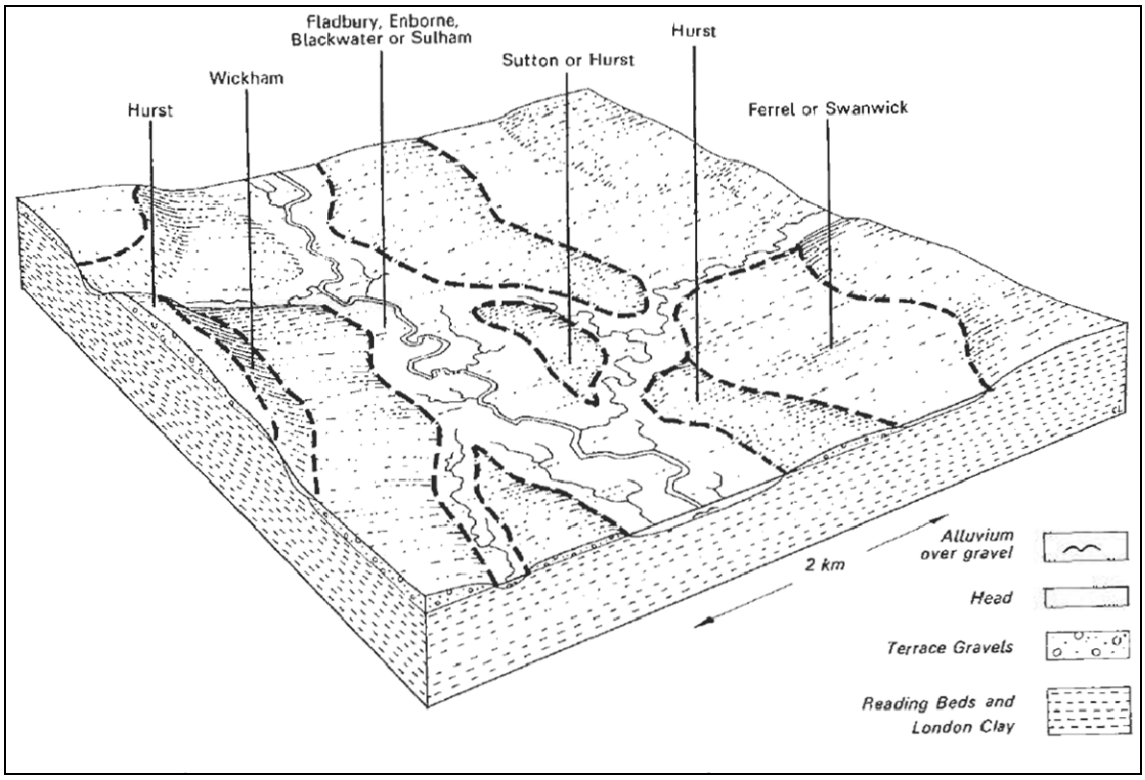


Figure 6 - Example block diagram showing the likely location of soil series within a soil association (from Jarvis et al., 1979)

Block diagrams were integral parts of the Regional Bulletins (e.g. Findlay and Soil Survey of England and Wales, 1984; Ragg and Soil Survey of England and Wales, 1984) which accompany the National Soil Map of England and Wales and were often used to describe the position of soil associations (groupings of soil series) within a landscape. These block diagrams less commonly describe the positions of soil series within the associations (Figure 6), and, unfortunately, in the UK, systematic soil survey has ceased and so other methods of determining and mapping soil parent material are required for use in areas without detailed soil mapping.

Description	Source	Scale	Techniques / Comments	Reference
Very general soil parent material map for region in Iran	Soil profile data with subsurface information	1:250,000	Geomorphic hierarchical downscaling method used to decompose the processes forming the landscape	Toomanian et al., 2006
Swedish survey of forest soils	Field mapping	ca. 1:10M	Interpolation using ordinary kriging on 23,100 points	Odell and Lofgren, 2006, Lundin, 2006
Adirondack Park, USA	Soil map	1:62,500	Retrospective classification of soil units to nine parent material classes	Roy et al., 1997
Barnstable County, Massachusetts, USA	Geology map	1:31,680	Reinterpretation from existing geology map	USDA 2002
Rhode Island, USA	Soil map	1:100,000	Translation from soil map to 17 parent material classes, combined with expert knowledge	RI USDA NRCS 2009
Soil Geographical Database of Eurasia	Soil map	1:1,000,000	Very general descriptions of geology	European Soil Bureau, 2001
Limited areas in England	Geology map	1:50,0000	1:1 translation of geological maps to parent material units map	Mayr et al. 2001
National Parent Material Map for Great Britain	Geology maps with supplementary information	1:50,000	Use of geological archives and field sheets to attribute geological data with descriptions of near surface lithology	Lawley, 2009
Irish Subsoils Database	Aerial photography, existing geological and soil mapping with 3,000 field samples.	1:250,000	manually created from the interpretation of stereo aerial photography, in conjunction with information derived from all existing published geological, parent material and soil mapping, supplemented by 3000 field samples	Fealy et al., 2009)
Prediction of parent material under a loess mantle, USA	SPOT image	10 m pixel	Cloud free SPOT image (only accounted for 16% of measured parent material variation over 31 km2)	Agbu and Olson (1992)

Table 1 - Comparison of a range of techniques for the mapping of parent material from the literature

Parent material maps may be derived from existing soil maps. For example, a new map of soil parent material (Roy et al., 1997) was derived from an existing soil map at 1:62,500 scale (USDA Soil Conservation Service, 1975) for the Adirondack Park in New York State. In this study the authors retrospectively classified the fifty soil map units (soil associations) in this area into nine parent material classes with particular comment on the pH and hydrological properties of the parent materials. A simplistic translation from soil to parent material has been done for Barnstable County, Massachusetts (USDA 2002). A 1:100,000 scale soil survey for Rhode Island has recently been reclassified as a parent material map (RI USDA NRCS 2009) with 17 well described classes. This map was derived to assist soil evaluations for septic systems as the surface geology map (RIGIS, 1989) only shows broad classes of quaternary deposits and did not provide the required level of detail. Expert knowledge of the soil scientist was used to add attribution to the existing classes (Turenne, 2009). Because all soil series in England and Wales have a defined soil parent material (Clayden and Hollis, 1984), for small disparate areas, it is possible to derive parent material maps or hydrogeological substrate maps (NSRI, 2009) from existing soil series maps.

Internationally, the 1:1,000,000 scale Soil Geographical Database of Eurasia (European Soil Bureau, 2001) has been attributed with parent material according to the European Soil Bureau (ESB) classification (Finke et al., 1998). This derived parent material map is necessarily general and of little use at a local scale. There remains, therefore a need to develop cost effective methods of mapping soil parent material in detail across the UK.

2.2 Creating parent material maps from other sources of information

While the creation of parent material maps from sources of information other than soil maps is not common, there have been a few examples, including the creation of parent material maps from geological mapping, geomorphic maps and water well logs. These are now discussed.

2.2.1 Geological Mapping

Mayr et al. (2001) investigated new methods of soil mapping in the UK. As part of this study they recognised the need for an improved parent material inputs and attempted to derive these from existing geological mapping by means of translation tables. Their study found that certain parent material units could be effectively predicted from the geological mapping while other units were less successful.

Developing from and expanding upon aspects of this research, the British Geological Survey (BGS) began a programme where one of the aims was to produce a national parent material map at 1:50,000 scale, which details the distribution of physiochemical properties of UK parent materials. BGS have just completed this work and have recently begun to report on the work done (Lawley and Smith, 2008; Lawley, 2009).

The BGS Soil Parent Material Map is based on the 1:50,000 scale digital geological map - DiGMapGB50 (British Geological Survey 2007). They used the BGS Lexicon (British Geological Survey, 2009) – an extensive database containing information on the age, lithology, location and thicknesses of geological units – to initially attribute the geological maps with likely parent material codes based on NSRI's classification (Clayden and Hollis, 1984). However, the most recent version (v4) of the map no longer provides a correlation to NSRI soil parent material classes, but instead has used the more lithological classification of the European Soil Bureau (ESB) (Finke et al., 1998). Lawley and Smith (2008) recognise that the BGS dataset has four key flaws, which make it, in places, a poor parent material map:

1) the emphasis on the bedrock at the expense of the superficial deposits,
2) the tendency to concentrate lithological descriptions on the un-weathered material
3) the lack of quantifiable descriptions of the rocks (subjectivity)
4) the inconsistent, patchwork nature of the sample and survey patterns.

To overcome as many of these issues as possible, BGS have undertaken a significant data mining exercise with their archive datasets, combined with terrain analysis, creating a new geological map with a greater emphasis on the near surface geology. They have also added surface deposits which were unpublished on the original geological maps.

While the actual BGS parent material map was not available for this research, the recent publication of the user guide (Lawley, 2009) provides a description of the dataset. While it may not provide a parent material map according to the English and Welsh precedent, there appear to be many useful components of this dataset for the creation of a parent material map, which may, in turn address some of the needs identified in Section 1.1. Eight key descriptive fields are incorporated in this dataset. The new BGS parent material classification is based upon the primary origin of the material (e.g. sedimentary – clastic), its dominant mineralogy (e.g. silica-clay) and its generalised texture (e.g. argillaceous) (Lawley, 2009). This information is supported by a large number of other fields describing a range of attributes of the substrate, including the age, hardness, engineering strength. Of particular interest is a field describing the variability of the spatial uniformity of the mapping unit. This field contains high, medium or low classes and represents an early attempt at mapping the confidence of the classification. The texture of soils found overlying the BGS parent material units have been allocated soil texture classes according to the NSRI soil texture system (Hodgson, 1997) using a mixture of measured samples and estimates. Three classes of likely soils are also described in this dataset: Heavy, Medium and Light.

In the future, BGS hope to improve their parent material map by the incorporation of quantifiable survey information, remote sensing data and traditional soil survey and profile information. Furthermore, they wish to include a form of confidence mapping with the map (Lawley and Smith, 2008).

Soil parent material	SPM Code	Actual sites	Predicted Sites	Agreed sites	Class value (ξ)
Lithoskeletal chalk	Bi	249	339	232	0.799
Marine alluvium	Eb	226	189	164	0.794
Lithoskeletal acid crystalline rock	Ba	77	65	50	0.707
Chalky drift	Eg	402	398	258	0.645
Loam with interbedded sandstone	Fp	38	57	29	0.623
Clay or soft mudstone	Fi	549	451	293	0.589
Drift with siliceous stones	Ei	1614	1284	833	0.579
Lithoskeletal ironstone	Bg	18	15	9	0.548
River alluvium	Ea	215	314	141	0.543
Lithoskeletal basic crystalline rock	Bb	36	33	18	0.522
Deep peat	Ac	182	87	58	0.461
Lithoskeletal sandstone (or slate)	Bl	276	353	123	0.394
Lithoskeletal limestone	Bh	226	138	57	0.323
Brownish clay	Fh	20	63	27	0.282
Stoneless drift	Ef	215	44	26	0.267
Lithoskeletal mudstone & sandstone or slate	Bm	315	174	61	0.261
Sand or soft sandstone	Fq	76	121	22	0.229
Clay with interbedded limestone	Fj	28	146	13	0.203
Calcareous colluvium	Ed	21	64	6	0.164
Lithoskeletal chert, quartzite or flint	Bf	7	6	1	0.154
Clay and sand	Fl	8	6	1	0.144
Loam (or soft sandstone, shale or siltstone)	Fm	163	92	12	0.098
Peat over lithoskeletal material	Aa	12	58	2	0.076
Lithoskeletal mudstone, shale or slate	Bj	12	476	4	0.053
Non-calcareous colluvium	Ee	16	102	1	0.025

Table 2 – Success rates in predicting soil parent materials from the NSI dataset by BGS
Note: Derived from Palmer et al., 2007. The class value metric is explained in section 4.5.1.2.

A review of version 0.1 of the BGS parent material map (Palmer et al., 2007) found that the age of the geological mapping underlying the digital geological map (some are as old as the 1880's) can affect the value and reliability of the resulting parent material map. It was found that easily identifiable parent materials, such as chalk and marine alluvium were very well predicted (Table 2). Conversely, certain parent material types which are defined by the presence or absence of particular material within the top 80 cm (Clayden and Hollis, 1984, Appendix 8) were very poorly predicted, as this information is rarely if ever recorded on geological maps. It may have been for this reason that the NSRI parent material classification is no longer used in the BGS dataset. Nevertheless, the most recent version of the BGS parent material map does contain two fields pertaining to gravel. The first describes whether or not a gravel can form by weathering from the geological unit. The second field describes the likely abundance of gravel in the parent material.

The BGS parent material map does not provide a parent material class according to the defined system for England and Wales. Nevertheless, this new dataset appears to offer additional attribution to that available from the BGS Lexicon (British Geological Survey, 2009) which may be valuable for the creation of more accurate parent material maps. A key requirement for parent material maps in the England and Wales is the development of more detailed soil maps. Therefore, concerns remain about the lack of defined links between the BGS parent material classes and soil series. This linkage to soil warrants future investigation, along with an investigation of the value of the descriptive and supporting fields units for the identification of traditional parent material types.

Geological mapping does not always provide full details on the distribution of surface deposits, to overcome this, the Irish National Subsoils Database was based on multiple sources of information. The result is a surface geology map, manually created from the interpretation of stereo aerial photography, in conjunction with information derived from all existing published geological, parent material and soil mapping, supplemented by 3000 field samples (Fealy et al., 2009).

Geological mapping is not lithological mapping. A typical map unit on a geological map may include sandstone, siltstone and mudstone. As a consequence, the spatial distribution of a particular lithology (such as siltstone) is unknown, and can often be very complex (e.g. Phillips and Marion, 2005). For geological purposes this chronostratigraphic framework has been satisfactory, as such beds are often thin and laterally discontinuous. For soil parent material map generation, however, such units may cause a problem as these three lithologies can give rise to very different soil types with different hydrological and ecological characteristics, even under the same conditions of formation.

Similar, but perhaps even more acute problems arise with the geological mapping of drift deposits. Glacial deposits and alluvium are often highly variable units, both spatially and in terms of lithological composition, for which little or no lithological descriptions have been made (Lawley and Smith, 2008). It is inevitable that such lithologically heterogeneous units will produce highly variable parent materials, so this presents a problem for the modelling of soil from these deposits. Mayr et al. (2001) have shown that the extent of most superficial deposits are underestimated. Furthermore, they identify issues with three Quaternary and Holocene deposits. Peat is not shown where it is less than a meter thick, colluvium is shown only locally and loess appears to be missing apart from in parts of East Anglia. These types of deposits can be thin (less than 1 or 2 meters) and blanket the landscape. While notes of these units may have been made on the field sheets and published maps by cartographic means such as stippling, these thin units have often been disregarded in the creation of the 1:50,000 scale digital geological map (British Geological Survey 2007). Furthermore, it is possible that soils may transition into a geological material that is not the dominant soil parent material (Wysocki et al., 2005), particularly if the parent material was a very thin deposit. Thus, superficial deposits remain problematic for the creation of a parent material map.

2.2.2 Unpublished geological and soil maps

Unpublished soil and geological maps exist, and contain additional information for areas which may not have detailed published maps. Maps such as these are being used by the BGS in their attempt to model the under-represented superficial deposits (Lawley and Smith, 2008). Likewise, in the redigitisation of the National Soil Map in 2000, detailed 'compilation sheets' were used as the basis for the linework due to the extra detail on these unpublished sheets.

2.2.3 Geomorphic maps

Geomorphic maps contain map units based on landform attributes and surface geology. These maps can provide both the relief and parent material components of Jenny's (1941) functional soil equation, and as such offer great potential for the identification and mapping of soil parent material units – in particular those related to surface processes (Figure 7).

In an area strongly influenced by drift and alluvial material, Bui et al. (1998) attempted to model soil distribution. Geomorphic maps existed for some regions within their study area, and where these existed, they were used, yet a significant alluvial plain had no such map. In order to model the highly variable soils likely to be found on the alluvial plain, they sought to reconstruct the environment of deposition and from this, predict the textural and mineralogical composition of the parent materials.

Because there is a great emphasis on surface processes and geology in geomorphic maps, for the purposes of modelling parent material geomorphic maps can be more suitable than geological maps, which tend to focus on bedrock geology (Palmer et al., 2007). For the majority of the Earth's surface, however, geomorphic maps do not exist, while lithostratigraphic maps do. This is certainly the case in the United Kingdom.

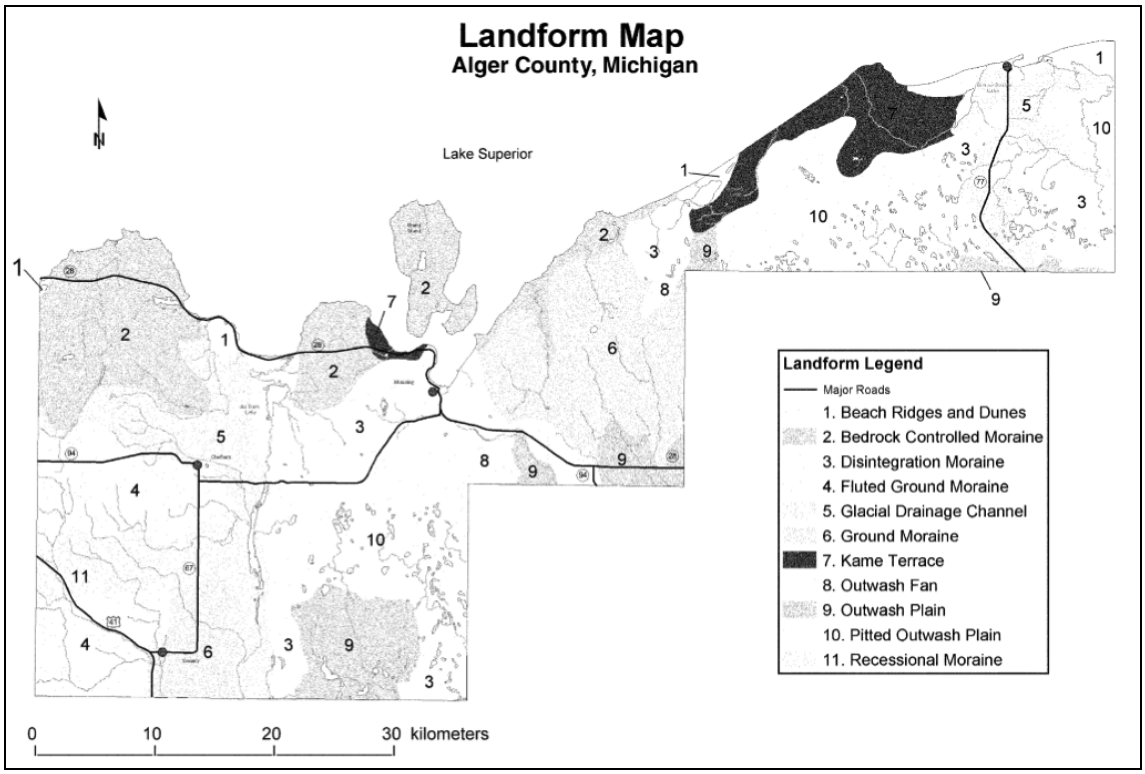


Figure 7 – An example geomorphic or landform map. From Wysocki et al, 2005.

2.2.4 Water well logs

To overcome problematic mapping of soil units formed on the Marlboro Clay regolith in the United States, Scott and Needelman (2007) used water well logs in a manner similar to bore holes to estimate thickness of this parent material unit. However they only achieved limited success at predicting outcrops of this parent material and conclude that, in their area, this form of parent material mapping is not suitable for assisting soil mapping.

2.2.5 Expert knowledge

Field surveyors build up a detailed and intuitive knowledge of the landscapes in which they work. A proportion of this expert knowledge is explicit, some is tacit. Where expert knowledge about soil parent material is explicitly stated, it can be useful when attempting to predict parent material. While no published studies have been found using expert knowledge to map or model soil parent material, attempts have been made to capture expert knowledge to aid soil mapping, although it has been noted that the use of expert systems to map soil properties needs further exploration (Scull et al., 2003).

While there are no readily available datasets for 'expert knowledge', certain knowledge has been formalised within databases such as NSRI's LandIS and the BGS Lexicon. In terms of soil parent material, what is represented in this category is the intuitive understanding of the relationships between geology, climate and relief in a landscape based on years of experience.

Work has been carried out attempting to glean from soil maps an understanding of soil surveyors mental rules (e.g. Bui et al., 1999) and yet there will always be more knowledge than can be captured from maps for use in parent material modelling. McKenzie and Ryan (1999) note how difficult it is to include intuitive mental models in explicitly defined models.

Interviews have been used by some to formalise expert knowledge of soil relationships into a suite of rules (Zhu, 1997; Qi and Zhu, 2003; Qi et al., 2006; Zhu et al., 1996) and there appears little reason to suppose why similar approaches cannot be applied to parent material. To avoid problems of surveyors subjectivity which can arise through interviews, Lagacherie and Holmes (1997) attempted to build expert knowledge rules from detailed soil maps. It has been noted that it is beneficial to gain as much understanding about an area as possible before mapping (Findlay, 1970). This also applies when attempting to extract and formalise expert knowledge in order to maximise the accuracy of the mapping or modelling.

The majority of soil surveyors in the UK, formerly employed by the Soil Survey of England and Wales (SSEW) have now retired. While there is potential to interview a very small number of soil surveyors to extract knowledge about relationships between parent materials and the landscape in familiar geographic regions, there appears to be a body of knowledge, that is published literature, which has not yet been queried for information on soil parent material. Indeed, little work has been done attempting to extract expert knowledge regarding soil or any of its properties from published literature, and this body of knowledge offers significant potential for building models of parent material.

The relationships between soil associations and landuses are described in the Regional Bulletins (e.g. Findlay, D. C. and Soil Survey of England and Wales, 1984), yet these offer little information on parent material or underlying geology due to the coarse scale of the National Soil Map (NSRI, 2008b) to which they refer. However, several key relationships are described in Soil Records (e.g. Colborne and Staines, 1987; Jones, 1983; Hollis, 1978; Reeve, 1976; Sturdy, 1971). These include the relationship between soil types and underlying geology, slope and elevation (Figure 8). Because the areas mapped in detail (Figure 3) were chosen as representative of the soil landscapes surrounding them, there may be opportunities to use this expert knowledge to predict parent material beyond the study area. If this information, captured in the literature, was systematically extracted and formalised into rule-sets or probability functions, there might be potential for the use of geological mapping or digital terrain models (DTMs) to map parent material using this extracted knowledge. This approach of extracting rules or probabilities from literature appears to be novel. Furthermore, descriptions of typical land uses and drainage patterns are sometimes described which may allow mapping of such units from aerial photography.

Table 3 Classification of Soil Series (Cont'd)

Group	Subgroup	Definition	Series
	Pelo-stagnogley soils	Clayey material passing to clay or soft mudstone	DALE
		Swelling-clayey material passing to clay or soft mudstone	DENCHWORTH
		Clayey drift with siliceous stones	HALLSWORTH
Alluvial gley soils	Typical alluvial gley soils	Medium silty river alluvium	CONWAY
		Clayey river alluvium	FLADBURY
	Pelo-calcareous alluvial gley soils	Clayey river alluvium	THAMES
Earthy peat soils	Earthy eutro-amorphous peat soils	Humified peat	ADVENTURERS'

The purpose of a soil survey is to describe, classify and map soils in order to aid decisions about land use and management. To do this on a national basis, soils are grouped into soil profile classes using the schemes of Avery (1980) and Clayden and Hollis (1984). Soil profiles are classified at four levels of generalization: major group, group, subgroup and series, according to their composition and the processes involved in their formation. Soil series have a limited and defined range of properties and horizons and are developed in lithologically similar material. Series names are taken from where they were first described or are extensive. Soil phases show the distribution of characteristics locally important but not differentiating at series level, for example slope, stoniness or topsoil texture.

The field to field pattern of soil variation is often too intricate to be mapped exactly even on large scale maps, so the distribution of soil series as shown on the map usually contains some included soils. In this survey fieldwork was carried out with frequent observations using a soil auger, at an average density of 30 per square kilometre.

Representative profile descriptions of soils previously unrecognized are included in the Appendix, while detailed descriptions of profiles of other

soils can be found in Soil Survey publications as indicated in the text. The profiles were described following the terminology of Hodgson (1976) and were sampled for laboratory characterization. Methods of laboratory analysis are given in Avery and Bascomb (1982).

Soils, geology and landscape

Soils on the Yeovil and Pennard Sands cover more than half the district and are strongly influenced by the lithology of the rocks beneath them. In both strata fine textured lower layers grade upwards to coarser sandy beds, forming heavier soils on the lower ground and lighter soils on upper slopes (Fig. 3). Soils on lower and middle beds of the Yeovil and Pennard Sands

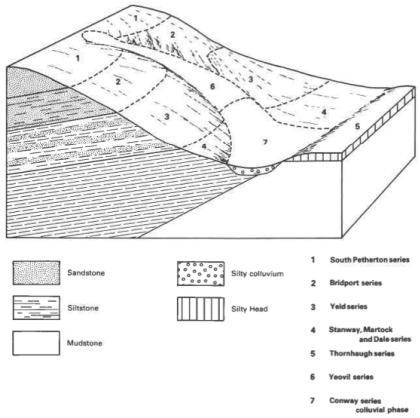


Fig. 3. Soil, geology and landscape relationships on the Yeovil Sands

are clayey (Dale series), medium silty over clayey (Martock series) and medium silty (Stanway and Yeld series). The medium silty soils are most extensive and cover half the outcrop. The grain size of the upper Yeovil and Pennard Sands ranges from coarse silt (median value 54 μm) to very fine sand (median value 72 μm) (Davies 1969). On these, South Petherton soils are most common over the coarsest, very fine sandy beds, and Bridport soils over sediments with a large content of coarse silt.

Silty colluvium or hillwash covers almost a tenth of the Yeovil and Pennard Sands outcrop. Light silty soils (Yeovil series) are most common. Thornhaugh soils developed in silty drift derived from the Yeovil and Pennard Sands occur in many valleys.

Middle and Upper Jurassic clays and mudstones, which underlie a fifth of the district, have smectitic clay mineralogy (Avery and Bullock 1977). Denchworth and Evesham soils on the Fuller's Earth or Forest Marble clays are strongly swelling. However Dale soils over the more silty, less calcareous Oxford Clay have smaller shrink-swell potential.

ABERFORD SERIES

These are calcareous medium loams (typical brown calcareous earths) over limestone at between 30 and 80 cm depth. They are on level or gentle slopes, most commonly over Junction Bed limestones but also over the Ham Hill Stone and Inferior Oolite. Aberford soils are most extensive around Stoke sub Hamdon. The brown, variably stony clay loam topsoils are usually calcareous when cultivated but under some long established grassland are non-calcareous. The stony, yellowish brown, weakly or moderately structured, calcareous clay loam subsoils overlie hard limestone which is often sandy, shelly or oolitic.

Brief profile description (full description Hartnup 1975, p.35)

0–25 cm Ap
Dark brown, slightly or moderately stony clay loam.
25–55 cm Bw
Brown or yellowish brown, slightly or moderately stony clay loam; moderate coarse subangular blocky structure; calcareous.
At 55 cm R
Limestone.

Soil water regime

The soils are well drained and permeable (Wetness Class I). Available water reserves vary with depth to rock, but where there is a 55 cm thickness of soil over rock, available water reserves are moderate (115 mm). Such soils are slightly droughty for the common cereal and root crops throughout the district, but for grass they are very droughty in the north and moderately droughty in the south.

Cultivation and cropping

There is ample time for cultivation in both autumn and spring and the chief limitation on land use is droughtiness, which varies with the depth of soil. Soils greater than average depth are well or moderately suited to cereals. Droughtiness combined with a slightly limited landwork period, makes the soils only marginally suited to sugar beet. When irrigated they are well suited to potatoes but are droughty without it. The land is suited to intensive grass production, having small poaching risk, although potential yield is limited by drought.

Associated soils

Similar but non-calcareous Waltham soils (Courtney and Findlay 1978) occur sporadically, particularly over the Inferior Oolite and limestone in the Junction Bed in the north-east. Light loamy non-calcareous Dinorben soils over limestone (Clayden and Hollis 1984) occur over Ham Hill Stone on Chiselborough, Chinnock and Hamdon hills and between Crewkerne and North Perrott. In places over the Junction Bed limestone around Lufton and Thorne, clayey, medium loamy over clayey and medium silty soils are found. Elmton soils form sporadically wherever the limestone occurs at less than 30 cm depth.

ADVENTURERS' SERIES

Adventurers' soils are deep, base-rich amorphous peat soils (earthy eutro-amorphous peat soils), affected by high groundwater during much of the year, which are found in two small patches on Coker Moor. They consist of more than 40 cm of black amorphous, greasy peat. When cultivated, topsoils have a fine blocky or granular structure while subsoils are weak blocky or prismatic. The peat overlies silty or clayey, greyish, mottled alluvium at depths between 50 cm and more than 1 metre.

Figure 8 - Extract from Colborne and Staines (1987) with potentially useful information highlighted

Note on colours: Yellow: parent material / geology; Green: general interest; Blue: landscape, slope; Red: specific soil series names; Purple: geographic location of outcrops.

2.2.6 Remote sensing and geophysics

Remote sensing and geophysics offer a suite of tools for characterising the surface and subsurface. These techniques offer the opportunity of defining map units or refining those predicted by other datasets or the methods discussed above. In particular, remote sensing adds the ability of a more quantitative approach of defining some key characteristics of the surface layers – including the particle size distribution of the topsoil and lithological and mineralogical characteristics of the near surface which may not be described in detail on existing geological mapping. Remote sensing techniques may also provide information on deeper layers – providing indications of water levels or soil depth which may in turn provide further information on the parent material.

Worral et al. (1999) present a helpful table describing various techniques and the depth they assess. An adapted and significantly expanded form of this table is presented in Table 3, including the applicability of the techniques to landscapes under the temperate climate found in the UK.

2.2.6.1 Aerial photographic interpretation

Aerial photography, both visible and infrared, has been shown to be a powerful predictor of lithology and surface geological features (Slaymaker, 2001; Gomez Valle et al., 1970) and has been used for many years as part of desk studies preceding field soil survey, relating vegetation and land use patterns in the photos to change in geology, soil type or water regime. Aerial photography has also been used in the mapping of economic near surface deposits (Dowling, 1966) and glacial limits (Svendsen et al., 2004).

Table 3 – Possible geophysical and remote sensing techniques for parent material mapping within the UK
Note: PM refers to parent material

Technique	Radiation	Sampling depth	Advantages	Limitations
Aerial Photography	Visible / Infrared	surface	Total UK coverage, inexpensive, high resolution, pattern recognition used to guide surveys	Great experience required to interpret photos, subjectivity, PM rarely exposed so must be inferred
EM Remote Sensing	Visible / Infrared	surface	Total UK coverage, inexpensive, quantitative, many spectral bands, selective use may identify texture / lithology	PM recognition limited by vegetation and cloud cover in the UK.
Radiometrics	Gamma rays	< 1 m	Good definition of mineral composition, lithological responses well understood	Very expensive, very limited coverage, strongest signal from top 20 cm.
Radar / GPR	Microwave	1-20 m	Can provide PM thickness as well as composition	Limited to site scale due to cost / time
EMI	VLF	1-5 m	Differentiate texture fractions and lithological units	Limited to site scale due to cost / time
AEM	VLF	10-100 m	3D mapping, particularly in more weathered terrain	Measures conductivity so arid environments preferred
Magnetics	Magnetic field	N/A	Low cost survey of deep geology	Limited applicability to surface regolith
Digital Elevation Models	N/A	N/A	Low cost, total UK coverage, quantitative identification of breaks in slope, floodplains, river terraces	Interpretation usually required in conjunction with other information. Artefacts occur and are difficult to remove

The major limitation of aerial photography or satellite imagery in mapping soil parent material is that the parent material is often obscured by the soil, unless revealed by erosion (Agbu and Olson, 1992). Thus in order to gain an understanding of the parent material, there is strong reliance on the interpretation of surface expressions of subsurface features, which requires considerable experience (Avery and Soil Survey of England and Wales., 1987).

2.2.6.2 Multispectral remote sensing images

Visible light and multispectral optical sensors have been the mainstay of remote sensing for many decades. Aerial photographs and images produced by sensors such as Landsat, once exclusive to government and researchers, have now become easily accessible by the general public. There has been a continual improvement in the resolution of the satellite sensors, both in terms of the number of spectral bands and the spatial resolution of discernable features on the ground.

Despite the many advances in sensor technology, remote sensing can only rarely be used for parent material differentiation in temperate climates as the energy reflected or emitted by vegetation often strongly masks that of the energy from the soil or the underlying rock. This limits the information which can be gathered in such ways.

In some situations, predominantly drier environments, there has been some success using multispectral remote sensing systems to obtain information about the soil and its properties. The requirements for this are generally cloud free images, bare soil (Peng et al., 2003) and dry environments (Odeh and McBratney, 2000), and so are typically unsuitable for use in the UK. In Illinois, Agbu and Olson (1992) constructed a model to predict parent material under a loess mantle using a cloud-free SPOT image with the majority of the fields having bare soil. However, even given these ideal conditions, the model variables only explained 16.2% of the measured parent material variation.

Some work has used multispectral remote sensing as a correlative for parent material to aid digital soil mapping. Sommer et al. (2003) used a rule based method with soil,

organisms, relief and parent material inputs derived from an airborne Daedalus-ATM remote sensing system, with a ground resolution of $1m^2$, to predict surface gravel content for precision agriculture in a complex area of fluvial sediments. This was carried out over a very small area ($0.1km^2$), and the use of such airborne data over larger areas would incur significant expense. Lagacherie et al (2008) successfully estimated clay and calcium carbonate contents from bare soil using an airborne hyperspectral system.

While some studies have used remote sensing in vegetated terrains to discriminate lithological units, these often include high resolution gamma radiometrics, field work and magnetic data as well (Schetselaar et al., 2000; Wilford, 1992). Nevertheless, future work has been proposed investigating the use of sensors such as Landsat and ASTER to improve near surface models (Lawley and Smith, 2008).

Multispectral remote sensing has long been used as a tool in geology to map both lithology (Won-In and Charusiri, 2001) and geological structures (Chatterjee, 2003), but the majority of these studies have been made at regional scales over remote areas where geological mapping is limited. To map a large area of Australian regolith, Laffan and Lees (2004) used Landsat data in conjunction with 40,000 drill cores, in an attempt to draw relationships between the datasets in order to predict the situation in unknown areas.

For areas where detail is not required, Odeh and McBratney (2000) found the use of a regression / kriging model on 1.1km resolution AVHRR images with DEMs to be effective at predicting the clay content of soil. Soil texture is strongly controlled by parent material. Once more, the images were acquired on cloudless days outside the growing season. Some techniques developed for remote sensing are also being used in proximal soil sensing to determine mineral composition using portable visible – near infrared spectrophotometers (Viscarra Rossel et al., 2009).

Electromagnetic waves in the thermal range have been used in remote sensing to discriminate between lithologies. Zhang et al. (2007) used the thermal capabilities of the

ASTER sensor to extract information on the lithology and mineralogy of surface geology in California. Similar approaches have been used in Australia (Hewson et al., 2005), Namibia (Gomez et al., 2005), and Iran (Moghtaderi et al., 2007). The vast majority of this exploratory work has been drawn from the mining and mineral extraction industries in arid and typically inaccessible regions with little previous geological mapping. Again, vegetation and cloud cover cause significant problems for such approaches.

2.2.6.3 Gamma radiometrics

Digital soil mapping has made use of gamma radiometric data, where available, as a correlative for parent material on its own or in conjunction with traditional geological maps (Mayr et al., 2001; Cook et al., 1996; Wilford et al., 1997; Koons et al., 1980; Dickson and Scott, 1990; McDonald and Pettifer, 1992). While such approaches show great potential, this type of quantitative radiometric data is not currently nationally available in the U.K. (Mayr et al., 2001), and where it does exist, it is very expensive.

Gamma radiometric survey is a common geophysical technique used to discern textural or mineralogical information about geology and soil (e.g. Sommer et al. 2003; Cook et al., 1996; McKenzie & Ryan, 1999; Ryan et al. 2000). Variations in the natural emissions of gamma radiation from rocks or their derivatives can be used to aid the mapping of soil parent material (Cook et al., 1996). It is common to measure the radiation in three bands, corresponding to Potassium (K) Uranium (U) and Thorium (Th). Concentration differences can be useful in differentiating lithologies or geological origins (Tzortzis et al., 2003).

Acidic rocks emit high levels of radiation in all three windows or bands, while mafic rocks and sand have low signals (Cook et al., 1996). Schetselaar et al (2000) found the glacio-fluvial deposits in their area rich in potassium, highlighting them on the radiometric survey. Likewise Ramli (1996) found low gamma readings in areas of drift and peat. Such patterns can be useful in delineating such parent materials which have

been found to be poorly predicted in other parent material maps (Mayr et al., 2001; Palmer et al., 2007).

Because gamma radiometry has been used by the mineral industry for some time, the responses of rock are well understood. Weathered material responses are less well known (Wilford et al., 1997) but there is now growing interest in this area (Lawley and Smith, 2008; Dickson and Scott, 1990; McDonald and Pettifer, 1992).

Wilford et al. (1997) found a good correlation between land and air measurements for K and Th, but not for U. They split gamma ray response into two categories; primary (pertaining to the lithology) and secondary (pertaining to alteration, weathering and pedogenesis). While the secondary sources are very complex, trends were possible to find.

During weathering, radioisotopes are released into the regolith, but, due to leaching and other processes, the overlying regolith will not necessarily have a similar signal to the fresh rock (Wilford et al., 1997). Leaching can deplete the K in the regolith, while the U and Th values may be elevated due to the presence of iron oxides or clays in the profile (Wilford et al., 1997; Koons et al., 1980; Dickson and Scott, 1990). This can be problematic as the gamma radiometrics give a good indication of the properties of the top 30 cm, with stronger emphasis on the top 15 cm. Such depths are often too shallow to accurately map true parent material.

Soil, water and vegetation can attenuate the gamma rays from reaching an airborne sensor (Wilford et al., 1997). Forests can strongly attenuate the rays, but this should not be problematic in England or Wales, where much of the countryside is open. Some researchers have used Landsat TM imagery to derive wetness to correct for this problem of vegetation (Lavreau and Fernandez-Alonso, 1991). Plant tissue contains negligible traces of U and Th, but may contain high amounts of K. Kogan et al (1969) show that vegetation emissions can contribute up to 15% of the recorded gamma radiation in the K band. (Cook et al., 1996; Wilford, 1992).

Wilford et al. (1997) suggest that gamma ray energy is reduced by 20% with an increase of 20% soil moisture content. Similarly, Cook et al. (1996) indicate that there is a loss of 1% per 1% increase in soil moisture. However, these interactions are complex. Because of this attenuation of the radiation, the consensus is that it is best to do a ground and aerial survey at the same time, and when the ground is as dry as possible.

Minty (1996) describes in more detail possible processing and correction procedures for gamma ray surveys. These remove the effects of cosmic rays, scattering and changes in the elevation of the aircraft. Gamma ray footprints can be large. Wilford et al. (1997) explain that 60% of the gamma radiation received at a height of 100 meters comes from a 120 m radius footprint on the ground. Thus, small features can be lost in the noise. Noise reduction methods are discussed by Dickson (2004).

Some limitations of gamma-ray surveys, outlined in Wilford et al. (1997) are as follows: Not all regolith units can be identified by their gamma-ray response, as different regoliths can have the same response. Other areas may have no gamma ray production. There is variation in gamma-ray abundance due to soil moisture, which changes over time. This is hard to separate from the regolith response. Finally, small scale features can be missed due to the wide spacing of flight lines and the large footprint. Conversely, benefits of gamma-ray surveys include the easy provision of information on surface geochemistry, the distribution of primary and secondary minerals (Caspari et al., 2006) and the style and distribution of weathering across the landscape.

Combined with aerial photographic interpretation, satellite imagery and elevation models, gamma radiometry can be used to discern weathered materials much more effectively than it can on its own. Unfortunately, the cost and lack of availability of this data at national coverage make it a promising, yet currently unsuitable parent material correlative for parent material mapping in England and Wales at this time.

2.2.6.4 Ground penetrating radar

Ground penetrating radar (GPR) uses microwaves to detect reflected signals from subsurface layers or structures. Gerber et al., (2007) used ground based GPR to successfully map the thickness and spatial distribution of periglacial slope deposits, which are a major soil parent material in the mountainous regions in Germany. While in the majority of studies using GPR, the antennas touch the physical surface, it is theoretically possible to use GPR from an airborne platform (Sen et al., 2003). However, this approach is in its infancy, therefore a labour and cost intensive approach is currently required for the identification of subsurface deposits. This tends to limits the use of this technology to site specific investigations (Freeland et al., 1998), and is not suitable for application across larger areas.

2.2.6.5 Electromagnetic induction, magnetic susceptibility, and very low frequency remote sensing

In some instances, and often at field scale, very low frequency (VLF) and electromagnetic induction (EMI) geophysics have been used with success to identify textural and lithological properties of the soil and parent material (e.g. Sommer et al., 2003; Cauvin-Cayet et al., 2001). Building an understanding of the origin of parent material sediments, Feng and Johnson (1995) isolated the magnetic susceptibility (MS) of the sand, silt and clay, finding different MS for the silt and sand fractions depending on the stratigraphic units of origin. James et al. (2003) used handheld EMI scanning techniques to determine boundaries of three soil classes over slightly different parent materials. When differentiating between three soil classes they achieved an agreement of 26% using Cohen's kappa statistic (Cohen, 1960). Reducing this to two classes resulted in an agreement of 62%. EMI in conjunction with electrical resistivity tomography (ERT) and ground sampling was used to locate the presence and depth of a gravelly parent material in a vineyard (Morari et al., 2009). The knowledge of soil and parent material MS properties has been used to assist landmine clearances (Hannam and Dearing, 2008) and to delineate hydric soils in the United States (Grimley et al., 2004). Nevertheless, these studies typically are field or lab-based and offer lesser scope as part of more cost-effective remote sensing platforms.

2.2.6.6 Airborne electromagnetics

Active airborne electromagnetic (AEM) surveys have been used to create three-dimensional maps of salt stores (Mullen et al., 2007), soil salinity (Macaulay and Mullen, 2007), and water resources (Dent, 2007). Recently, AEM has been proposed as a method of acquiring data on regolith materials at depth (Worrall et al., 1999). While this work produced some interesting results regarding the character of the regolith, for example the delineation of previously unknown paleochannels, the applicability of this technique to parent material mapping in the UK is unknown. This project was carried out in an arid cratonic region of Australia, where regolith weathering can reach hundreds of meters, therefore more work needs to be done examining the potential application of these surveys to the much thinner and younger regolith layers of the UK and under the influence of a temperate climate.

2.2.6.7 Magnetic surveys

Magnetic anomaly surveys have been used by geophysicists for some time, and are now being adopted by a wider community of users. Magnetic surveys, often taken in conjunction with potential gravity surveys, are very useful at discerning structural geology, faults (Benson and Hash, 1998) and igneous intrusive bodies (Maes et al., 2007) where there is a sharp discontinuity between rock types. Aeromagnetic surveys recognise magnetic anomalies which often lie at hundreds of meters depth (Galdeano et al., 2001). While geologists have used magnetic anomaly surveys for many years to find deeper bodies, their practical use for surface deposits or the mapping of soil parent material is limited.

2.2.6.8 Digital elevation models

Digital elevation models (DEM) and digital terrain models (DTM) can reflect the underlying geological structures by the way that these are expressed in the landscape. Breaks in slope or linear features may reflect changes in the surface geology. Sinowski and Auerswald (1999) used a DEM in conjunction with soil property measurements to map the boundary depth between two soil parent materials – quaternary and tertiary sediments. Stoorvogel et al. (2009) delineated three main geomorphological units: plateau, slopes and valleys using a SRTM (Shuttle Radar Topography Mission) DEM.

As the form of the landscape is predominantly controlled by the underlying rocks, certain lithological information may be inferred from the terrain surface. In a study examining the effect of slope position on soils, Agbenin and Tiessen (1995) note that the occurrence of multiple geological formations on a single slope significantly complicate the resulting soils.

Other studies have used the link between parent material and terrain to add information to their models. For soil modelling purposes, McKensie and Ryan (1999) subdivide the parent material into three components; substrate, aeolian accession (modifies the influence of the substrate) and erosion and deposition (influences soil depth). For the latter two, terrain attributes, in conjunction with geophysical remote sensing are given as potential environmental correlatives. Working with the relationship between geomorphology and geology, Krol et al. (2004) use an ontological approach for data integration across cartographic scales. They divide the landscape into seven geomorphic units and five complex lithological classes. The combination of these gives them 39 ontological classes which they then use to classify the landscape.

One of the main advantages of the use of digital elevation models as an environmental correlative is the consistent coverage and resulting quantitative data. Additionally, as DEMs can be derived from space and airborne platforms, using a variety of techniques (e.g. synthetic aperture radar or photogrammetric methods) they are relatively inexpensive to acquire.

2.3 Summary

There is a growing need for parent material maps, yet relatively few have been created. Where parent material maps have been made, there has been a strong reliance on existing soil or geological mapping, and the translation of these maps to parent material maps. There are limited detailed soil maps, but the whole of the UK has 1:50,000 scale geological mapping. Nevertheless, weaknesses have been shown with direct translations of geological maps to parent material maps. In particular superficial parent materials can be under-represented and chronostratigraphic units may give rise to multiple soil parent materials. The creation of parent material maps from existing geological maps should be tested and compared with map creation using other methods.

The potential for creating parent material maps for the UK from sources of information additional to geological maps should be investigated. There exist a number of techniques from related fields that may be applicable to mapping soil parent materials. However, within the context of the UK, many of the remote sensing approaches which have been examined offer little scope for providing significant additional information on the soil parent material. This is either because of the temperate climate with extensive vegetation cover, or due to a lack of availability of suitable geophysical data.

There is a wealth of information in the published literature relating soil types (and therefore their parent materials) to geological formations. The literature will be examined to determine the level of expert knowledge which may be extracted, and the use to which this might be put for modelling parent material. Furthermore, investigations will attempt to characterise the relationship between parent materials and the landscape, and to ascertain the relationships between these factors as recorded in published literature. The use of existing reconnaissance-scale soil survey maps will be examined to determine the potential for the use of these in guiding or enhancing parent material models in England and Wales. Data mining techniques will be investigated for comparison with maps produced by expert knowledge. Given the complex nature of the

parent material - geology relationship, multiple class membership will be investigated as a means of conveying parent material information.

In this study, a number of study areas will be selected, covering a range of landscape typical of those found in the lowland regions of England. Investigations will be undertaken to determine the value of parent material maps resulting from different data inputs and methodologies. Recommendations and conclusions will be drawn to guide future parent material mapping exercises in the UK.

3 STUDY AREA SELECTION AND DESCRIPTIONS

Three study areas within England were chosen to develop and assess new methods of creating soil parent material maps. The reasoning behind the choice of the areas is presented in this chapter. Descriptions are provided of the geology, landscape, parent material diversity and soils for each area. Reference parent material maps are provided for each study area.

There is no one definitive landscape for England and Wales. Each region has its own geological, geomorphological and cultural history. There are, therefore, a wide range of landforms and landscape characteristics across the two countries, and it is likely that different regions will give results of varying quality to certain approaches of modelling soil parent material. There is not scope within this research to investigate all landscapes in England and Wales, so a choice was made as to which areas would be appropriate to study.

Three study areas covering different landscapes were chosen. When making the selection of study areas, a number of factors were considered, including the existence of detailed soil mapping for reference purposes, geological and geomorphological history of the study areas, available datasets, and the extent of similar landscapes to the study areas. The similarity of landscapes was determined with the Soilscapes dataset (NSRI, 2008c) which groups landscapes on the basis of similar soils and ecological characteristics. The locations of the three study areas; Worksop, Needwood Forest and Yeovil are shown in Figure 9, along with the extent of soils similar to those found within these areas.

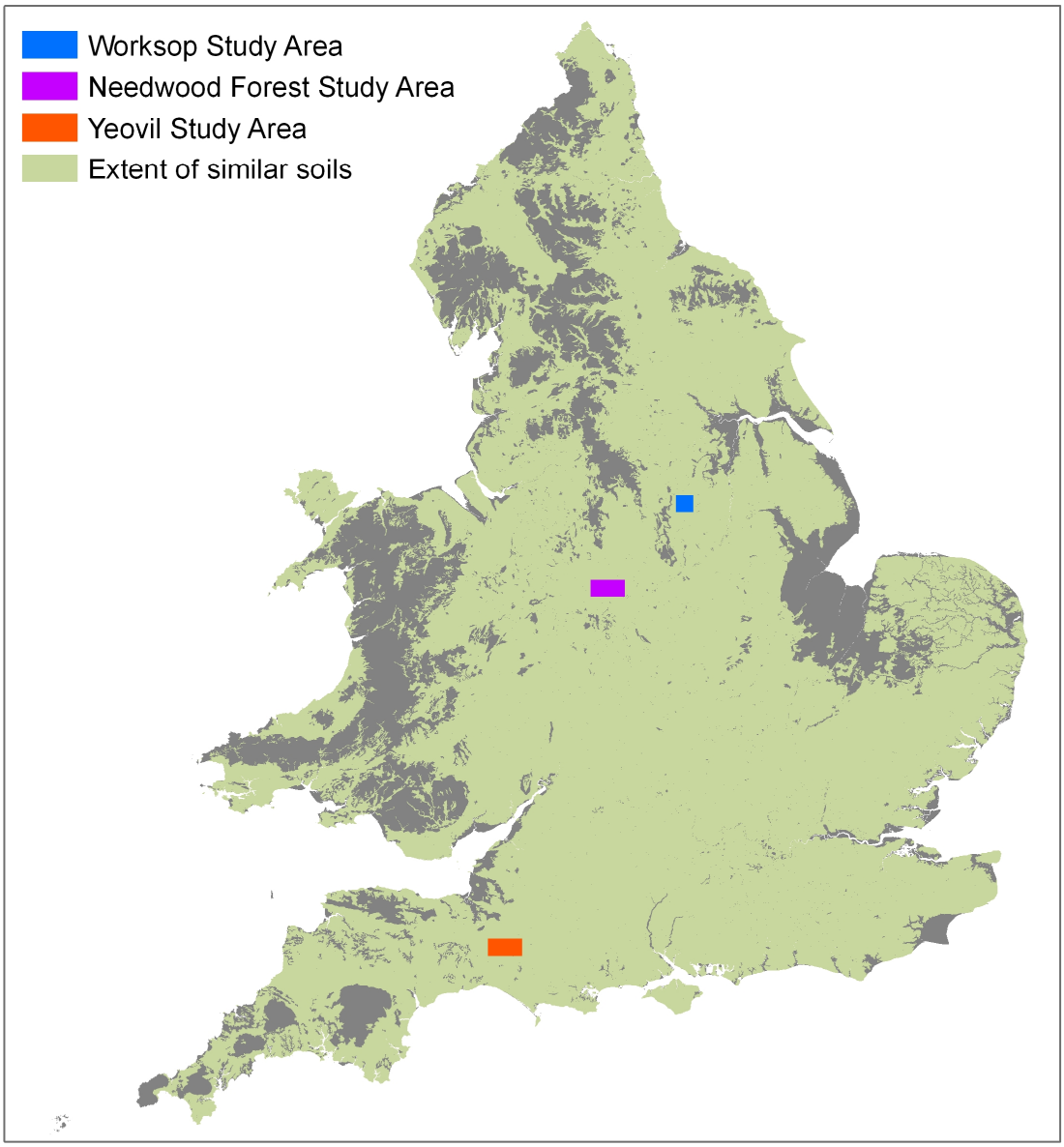


Figure 9 - The location of the three study areas and extent of similar soils within England and Wales.

Note: The extent of similar soils was derived from NATMAPsoilscapes, a simplified version of the National Soil Map. Soilscape units found within the three study areas were identified and their full extents are shown in green. Grey areas are typically upland peat soils or marine alluvium.

3.1 Geological diversity and soil landscapes

The bedrock geology in all three study areas is predominantly sedimentary in origin. The modelling of soil parent materials over a more diverse range of lithologies with the clear chemical and mineralogical distinctions usually found between igneous, metamorphic and sedimentary rocks would have produced maps with more easily distinguishable classes. However, most of the areas for which detailed soils information is frequently requested for agronomic, ecological, town planning or development purposes occur in the lowland regions of England and Wales. These are areas dominated by sedimentary deposits and it was felt that it would be more appropriate to develop methodologies in these regions where the differences between units are more subtle. Particularly desirable would be study areas that differed from one another in the extent of glacial superficial deposits, as the inconsistency in the ways in which these are mapped has previously caused problems in soil parent material modelling (Palmer et al., 2007).

Soils are often linked intimately with the underlying superficial deposits. Compared with most bedrocks, superficial deposits are less consolidated and so the processes of pedogenesis proceeds at a greater rate on these softer parent materials. Additionally a lower level of physical and chemical weathering is required for their incorporation into the solum.

The majority of landscapes in the United Kingdom have been affected by Quaternary glaciations. The effects of these include the removal of older soils and the deposition of extensive re-worked geological material in the form of glacial till and head. It is from these superficial deposits that much of the soil has formed. In comparison with, for example, the deep soils of the Bago region in Australia (McKenzie and Ryan, 1999) unaffected by glaciation for millions of years, the UK soils are thin and immature, most having been formed within the last 10,000 years.

Each glacial period resulted in reworking and reshaping of the landscape and the surface geology. Due to inconsistent historical mapping of Quaternary deposits, the extent and

composition of these deposits remains unclear. However, the strong relationship between these deposits and the parent material requires that these problematic deposits are considered.

The extent of superficial deposits and the quality of the geological mapping will affect the value of the resulting maps. Because of this, three areas of lowland England, each with a different level of influence from superficial deposits were chosen as test areas (Table 4).

The Yeovil sheet lies beyond the southern edge of all the major Pleistocene glaciations (Devensian, Anglian and the debated Wolstonian glaciations), while Needwood Forest is strongly affected by all three glaciations. The Worksop area is affected by the older Anglian glaciation and the Wolstonian glaciation (Shotton et al., 1993). All areas contain recent alluvial deposits.

3.2 Scale and quality of detailed soil maps

It was essential that the study areas had high quality, detailed soil information from which to train and develop models of soil parent material. NSRI holds a national soil map at a scale of 1:250,000 and more detailed published soil maps at a variety of scales ranging from 1:25,000 to 1:100,000. The geological data that was used in the research was at a scale of 1:50,000. Bearing this in mind, it may be logical to develop models using detailed soil data at the same scale, so as to avoid scale-related issues. However, the study areas that were chosen only have detailed soil series mapped at 1:25,000 scale.

The reason for using this larger scale data was a pragmatic one. There are only five 1:50,000 scale soil maps in England and Wales and much of these maps cover urban or coastal areas. Conversely, there are over 100 1:25,000 scale maps covering most of the notable soil landscapes of England and Wales. If this research was to produce results which could be at a later date expanded across the two countries, it was more sensible that the most extensive detailed soil maps were used.

Because of the subjective nature in which the soil maps were constructed, it was anticipated that there would be variation in mapping and reporting styles between soil surveyors. Each of the three study areas was mapped by different surveyors. It was felt that this was appropriate in order to develop methodologies which might be transferable between different regions and surveyors.

Table 4 - A comparison of the three study areas
Note: Different datasets provided different indications of the extent of superficial deposits, depending on the classifications used. This variation is presented in this table.

	Worksop (100 km^2)	Needwood Forest (200 km^2)	Yeovil (200 km^2)
1:25,000 Soils mapping	Published 1976 9 parent materials	Published 1983 11 parent materials	Published 1987 17 parent materials
1:250,000 National Soil Map	11 units Mapped 1980s	21 units Mapped 1980s	20 units Mapped 1980s
1:50,000 Geology mapping	15 units Four sheets, mapped between 1966 and 1974	15 units One sheet published 1982	26 units One sheet published 1973
Superficial geology /drift cover	6% (geology map) 26% (soils in thick drift) 46% (soils in thin and thick drift)	54% (geology map) 65 % (soils in thick drift) 83% (soils in thin and thick drift)	13% (geology map) 20% (soils in thick drift) 81% (soils in thin and thick drift)
Urban cover (soil map)	20%	5%	11%
Elevation range	30 – 160 m OD	50 – 150 m OD	5 – 250 m (most below 120 m) OD

3.3 The Worksop study area

The Worksop study area (Ordnance Survey map sheet SK57) covers 100 km^2 and lies on the borders of Nottinghamshire, Derbyshire and South Yorkshire. Although this area was affected by the Anglian and Wolstonian glaciations, it was unaffected by the Devensian glaciation, superficial deposits are not extensive in this region, and soils tend to be influenced by the underlying Permo-Triassic sedimentary bedrock of limestone, mudstone and sandstone. The 1:25,000 scale soil map for this area (Reeve, 1976) was converted into reference parent material maps (Figure 10) using the methods described in Chapter 4. Approximately 20% of the Worksop area is covered by urban conurbation or undifferentiated soils, for which no parent material information is available.

The Worksop area is a lowland region with an elevation range between 29 and 161m O.D. Relief is gently undulating, with the land rising to the west up the dip slope of the Magnesian (dolomitic) Limestone (Reeve, 1976). High ground to the east is formed by the escarpment of Bunter Sandstone Pebble Beds, which overlie older sandstones and marls. Drainage is predominantly to the east, towards the River Trent, by the means of three main rivers (Reeve, 1976).

The geological deposits were mapped between 1966 and 1974 on BGS sheets East Retford, Chesterfield, Ollerton and Sheffield. The superficial and bedrock units were combined to create a surface geology map which appears in Appendix 1. The surface geology of the Worksop area is dominated by the bedrock with some distinct lithological variation. Superficial deposits are present, but of limited importance to the soil parent material character in this area.

According to the geological maps, superficial deposits cover approximately 6km^2 of the 100km^2 area. The soil map differentiates between “thin drift” and “thick drift” (Appendix 2). It indicates that 46 km^2 of superficial material may be present in this area if “thin drift” (15 to 80 cm thickness) is considered, or that 26 km^2 if only “thick drift”

is considered. This uncertainty highlights an issue to be dealt with in this research; that of loosely-defined terminology and differences in mapping priorities of different surveys.

The soils in the Worksop study area were mapped at 1:25,000 scale around 1975, at roughly the same time as when the geological survey was being made. The linework from the soil map has subsequently been simplified for the 1:250,000 scale National Soil Map. The soils in this area are a mixture of loamy and sandy soils. To the west, the soils are loamy over the limestone and marl, while the sandy and pebbly soils overlie the Bunter sandstone pebble beds.

The parent material distribution in the Worksop area is relatively simple, with only nine units recorded by the detailed reference map (Figure 10) The western side of the area is dominated by limestone units (BhB1 and BhB2). The middle of the study area is a mixture of soils in thin drift, predominantly clay or soft mudstone (FiF1 and FiF2) with some sandy parent materials (FqF1). The eastern side of the area is heavily influenced by the underlying Bunter Sandstone Pebble Beds which, with reworking, have produced extensive drift with siliceous stones (EiE1) and some more pure sandstone units (BoB2). The other parent materials (alluvium and stoneless drift) are of limited extent and influence in this area.

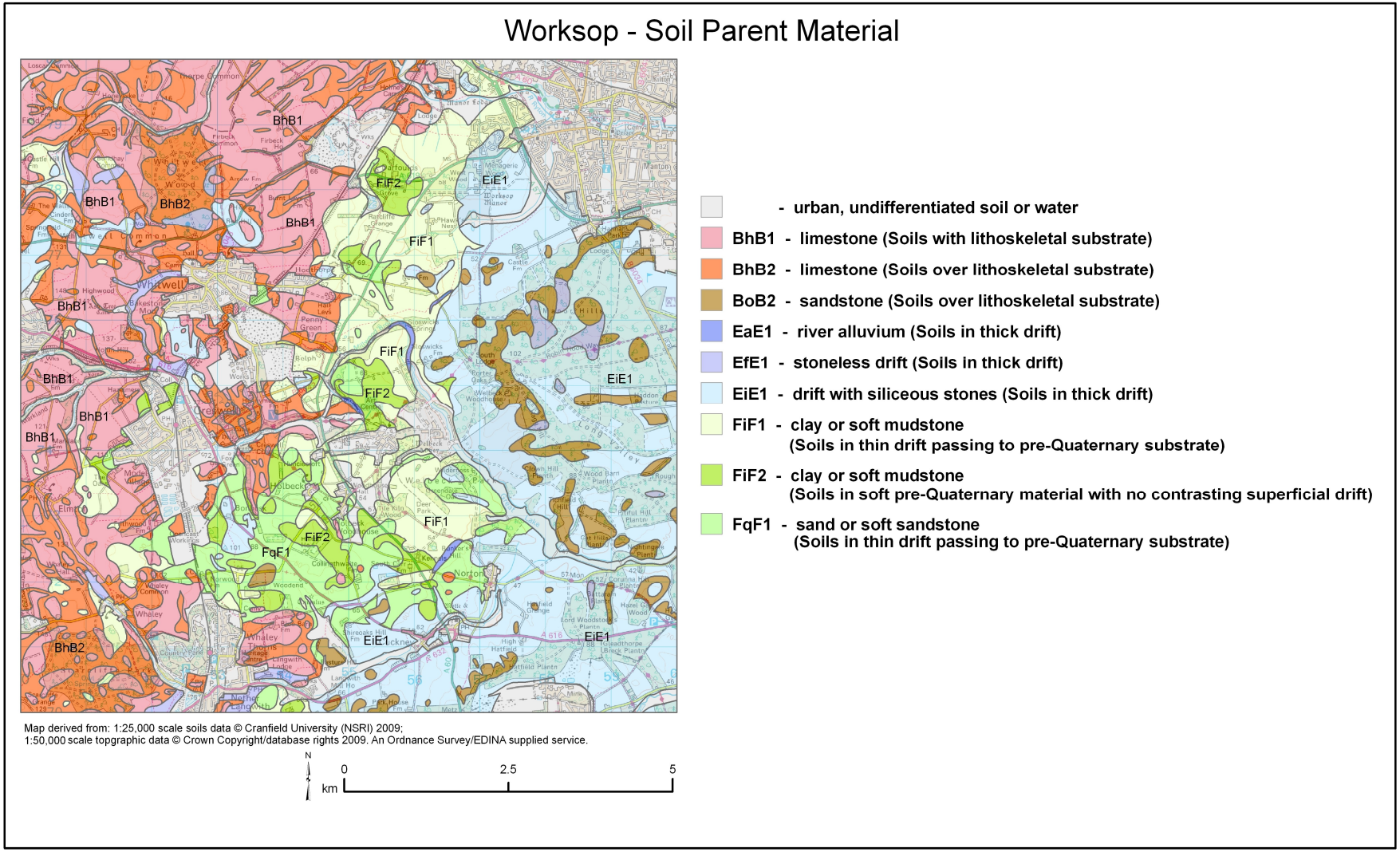


Figure 10 - Soil parent materials of the Worksop area (NSRI PARLITH classification)

3.4 The Needwood Forest study area

The Needwood Forest study area (Ordnance Survey map sheets SK02/12) covers 200km^2, of which 5% is covered by urban development or water. The Needwood Forest region has been affected by glaciations coming from the north, north-west and, during the Wolstonian, the north-east. This study area lies on the southern edge of the Devensian glaciation, and is dominated by till, morainic, glacio-fluvial and alluvial superficial deposits overlying otherwise extensive marls and shales of Triassic age (Jones, 1983).

Elevation ranges from around 50 m to over 150 m O.D.. Drainage is to the south by means of seven main river systems. For convenience, Jones (1983) divided up the region into landscape units; plateau, uplands, interfluves, valleys and a distinctive scarp region, characterised by steep slopes.

The geology in this study area is dominated by a diverse range of superficial deposits, yet the underlying bedrock marls and mudstones have an important part to play in the character of the area, both in terms of in-situ and glacially reworked material.

According to the geology map, 54% of this area is covered in superficial deposits. This is lower than figures obtained from the soil map for soils in drift (83%) or even just soils in thick drift (65%) This discrepancy results from the differing definitions of superficial deposits and mapping priorities of the two surveys, and thin drift deposits are particularly underestimated by the geological mapping.

Because of the extensive nature of superficial deposits in the region, bedrock less commonly forms the soil parent material in this area. The underlying bedrock geology is relatively simple, dominated by marly, mudstone deposits. The south-western corner of the map reveals the oldest rocks in this area – the pebbly Bunter sandstone, which also occurs in the Worksop study area. It is likely that, through reworking, both these deposits have contributed to the nature of the glacial drift which overlies them.

Because of the strong influence of glaciers on this region, superficial deposits comprised of local and erratic debris, dominate the surface geology.

The soils of the Needwood Forest area were mapped at 1:25,000 scale in the early 1980s (Jones, 1983), and this map, translated using the methods described in Chapter 4, forms the basis for the reference parent material map for this area (Figure 11). The linework from the detailed soil map was simplified for use in the later National Soil Map.

The parent material distribution in the Needwood Forest area is considerably more complex than in the Worksop area (Figure 11). Glaciation has created a patchwork of eleven different parent materials across the area with abundant drift deposits, both stoneless (EfE1) and with siliceous stones (EiE1). Clay or soft mudstone (FiF1 and FiF2) and be found in the southwest and northern areas. Alluvium (EiE1) and gravels (DbD1) are located in the river valleys and there is a small area of sphagnum peat (AaA3) in the northwest. A small area of soft shale or siltstone (FqF1) can be found in the southwest corner.

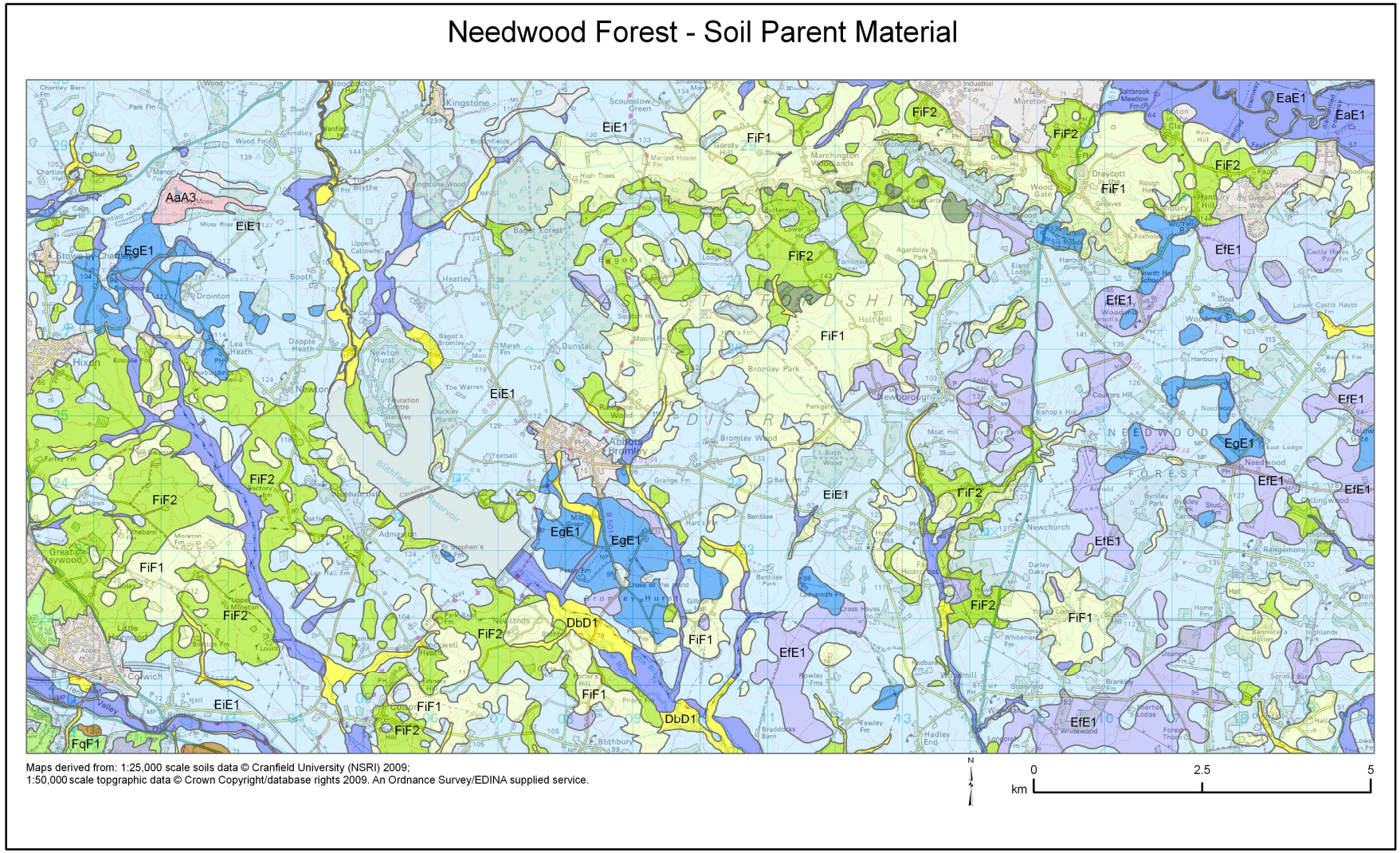
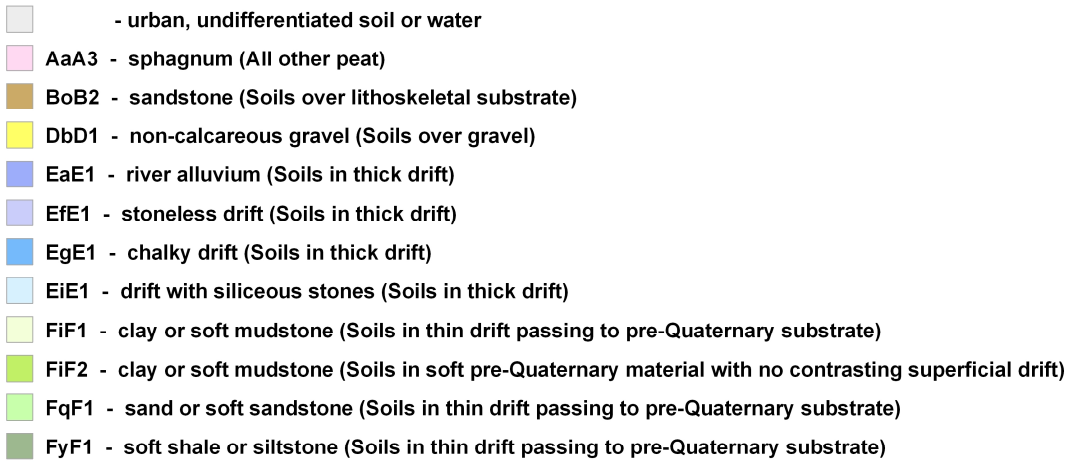


Figure 11 - Soil parent materials of the Needwood Forest area (NSRI PARLITH classification)

Soil Parent Material (Needwood Forest)



3.5 The Yeovil study area

The Yeovil study area (Ordnance Survey map sheets ST41/51) is the southernmost study area and covers 200km^2, with 11% of this area covered in urban development. Unlike Worksop and Needwood Forest, this area is unaffected by the major Pleistocene glaciations, and the surface geology in the Yeovil study area is predominately Jurassic bedrock consisting of sandstone, limestone and mudstone. Younger, Cretaceous chalk outcrops in the south-west of the region.

While the elevation of the land ranges from 6 m to 247 m O.D., most lies beneath 120 m. There are two main river systems flowing northwards with dissection of the land by tributaries increasing towards the south. The topography reflects the underlying geological structure with cuestas trending east-west with north facing scarps found throughout the region (Colborne & Staines, 1987).

The superficial deposits in the Yeovil area are derived from a mixture of alluvial and periglacial origin, but without the influence of glacial-related deposits which are so prevalent in the Needwood Forest study area. Once again, the superficial deposits are reported to be less extensive on the geology map (13%) than the soil map for both all drift (81%) and thick drift (20%). The majority of the superficial deposits have lithological characteristics indicative of their source material and the distinction between the bedrock and superficial deposit is clear. Elsewhere, the drift is harder to distinguish from the bedrock, as the drift material is sourced from the Yeovil Sands (Colborne & Staines, 1987) which are also found within this area. In some areas, colluvium can be several meters thick. Head is widespread, particularly in the south and gravel terraces are associated with both main river systems.

As the detailed soil mapping of this area (Colborne & Staines, 1987) was undertaken after the mapping of the National Soil Map, the linework for this detailed map was not used in the National Soil Map. This study area is therefore a particularly useful area for assessing the use of the National Soil Map as a predictor of parent material in regions

not mapped in detail. The detailed soil map was used to create the reference parent material map for this area (Figure 12) using the methods described in Chapter 4.

The Yeovil area has the most diverse parent material composition of the three study areas examined in this research, with sixteen distinct parent materials (Figure 12). Only the general pattern is described here. The north of the area is the least complex, and is dominated by soft shale or siltstone (FyF1). Moving south, loam or soft siltstone (FuF1) becomes more prevalent and then clay or soft mudstone (FiF1). Extensive dissection of the southern areas by the tributaries leads to abundant alluvium (EaE1) and some local stoneless drift (EfE1) or drift with siliceous stones (EiE1). Other small valleys are filled with deep colluvium (EeE1). Limestone parent materials (BhB1, BhB2) occur sporadically across the landscape where revealed by erosion. Further maps of the geology, soil and slope for each area are provided in Appendix 1.

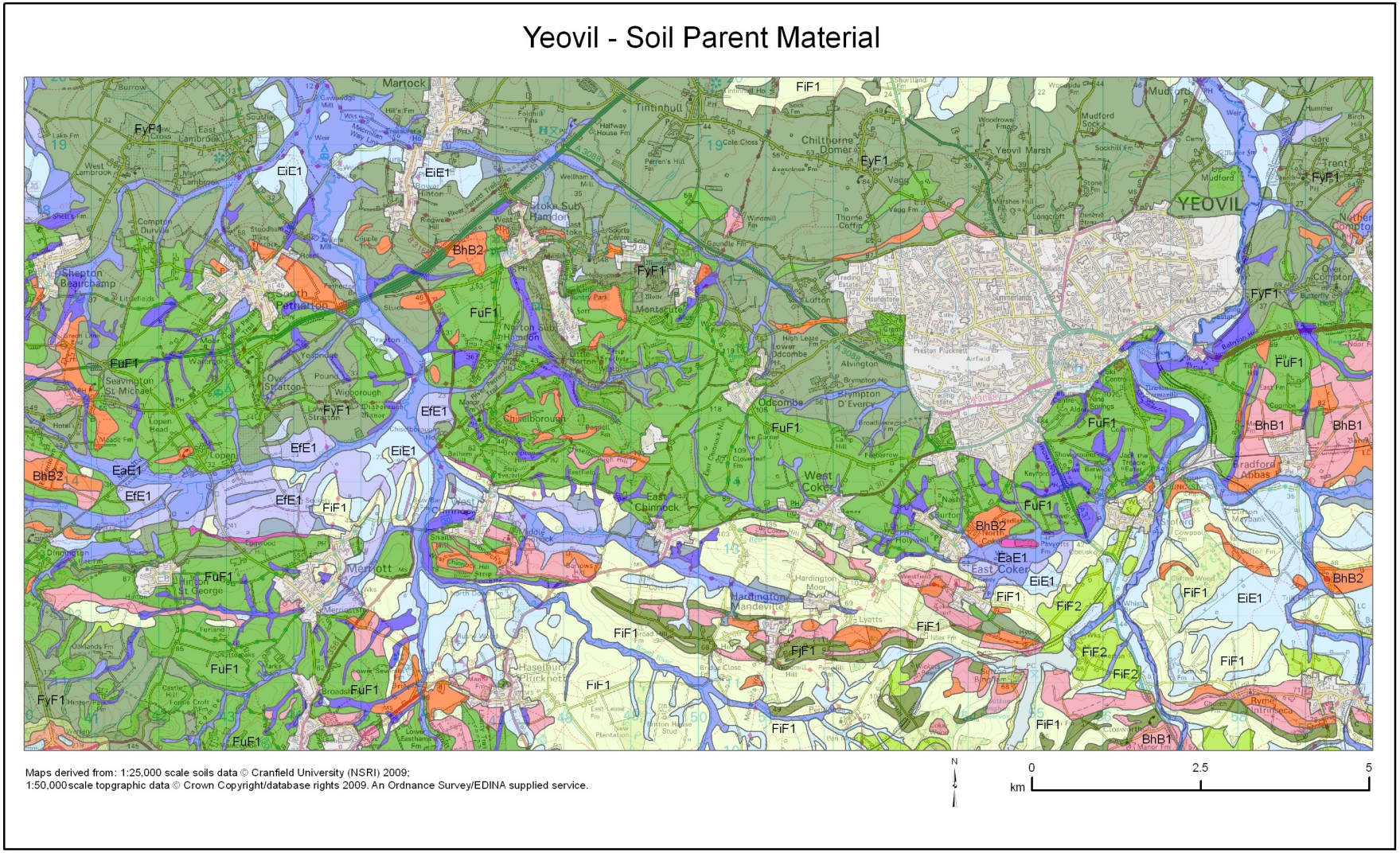
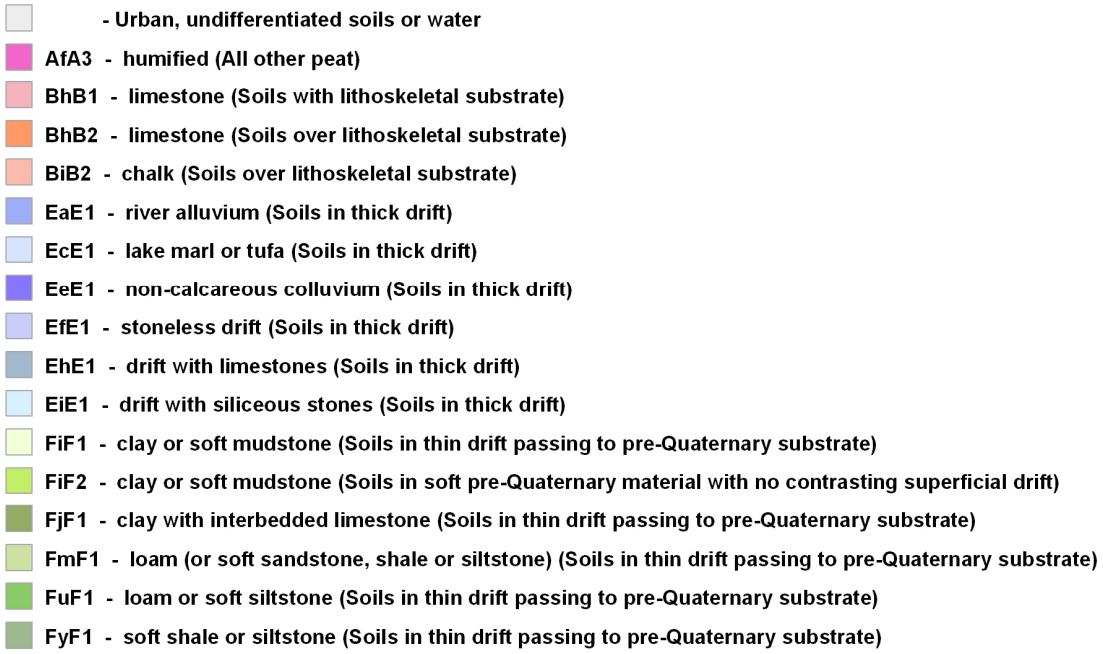


Figure 12 - Soil parent materials of the Yeovil area (NSRI PARLITH classification)

Soil Parent Material (Yeovil)



4 DATA, MODELS, METHODS AND METRICS

This chapter describes the data layers used within this research. The expert knowledge, data mining and combined methodologies use a model to combine the probabilities of occurrence of parent material classes, from multiple evidence layer inputs. This model, its inputs and outputs are described. The qualities of a useful parent material class and map are defined. Current map accuracy metrics are found to not wholly characterise the stated desirable qualities. Therefore these definitions are used to create new metrics of map value (ψ_3) and class value (ξ) for parent material maps. Supplementary metrics are also discussed as are the standard methods of result presentation and map analysis.

4.1 Data layers and preparation

Four main spatial datasets were used to create and assess the value of parent material maps. These were:

- the 1:25,000 scale reference detailed soil parent material maps
- 1:50,000 scale BGS bedrock and superficial geology (GEOLOGY)
- the 1:250,000 scale NSRI National Soil Map (SOIL)
- slope class map, derived from NextMap 5 m DTM (SLOPE)

The preparation of these data layers is described below.

4.1.1 Reference parent material maps

In order to assess the accuracy of the modelled parent material maps, two reference parent material maps for each study area were created using national and international parent material classifications (see section 5.2). These were the National Soil Resources Institute (NSRI) classification and the European Soil Bureau (ESB) classification.

Due to changes in soil classifications and nomenclature, certain soil series within the study areas have been combined with other series or renamed in accordance with the modern classification. Using the LandIS database, modern correlatives for the historic series names were found, and the rationalised (modern) name used. Undifferentiated soils, such as "undifferentiated bottomland soil" or "gorge soils" were ignored in these analyses and tests, as there is no defined parent material for such soils. In the case of joint or complex units, these were assigned to the dominant soil series in the unit. As all phases of the same soil series, for example, "stony phase" or "shallow phase" will have the same parent material, all phases were treated as the standard soil series in these analyses. These translations are described in Digital Appendix 1. Because multiple soil series share the same parent material, the resulting parent material map units are broader than those units describing soil series. This has also been previously noted by Wysocki et al., (2005) and Clayden and Hollis (1984).

The NSRI maps were derived from existing detailed 1:25,000 scale soil maps for the study areas, and translated to NSRI parent material classes using relationships defined in Clayden and Hollis (1984) (Figure 13). The ESB reference map was derived from the NSRI parent material map using an NSRI to ESB translation table created in this research (Figure 13). Areas where no parent material information was present (urban areas, lakes, undifferentiated soils etc.) were excluded from these maps and omitted from all analyses.

The detailed soil maps were used as the basis for the reference maps as the NSRI parent material is essential to the definition of a soil series (Clayden & Hollis, 1984), so it was assumed that the parent material would be correctly identified at the scale mapped.

Nevertheless, as soils are spatially variable, it is inevitable that not all the delineated area will be the identified soil series (Avery, 1987) and as a result, not all the translated area will be the defined parent material. There will be minor variations within the mapped soils and their parent materials. Indeed, when Sturdy (1971) investigated the homogeneity of the mapping unit on a 1:25,000 scale soil map in Essex, he found that while the majority of profiles within defined units were correct (typically 60 – 75%) there were some very heterogeneous units with a little as 47% of the soils identified in the unit corresponding to the named soil series for that unit. Furthermore, in complex soil units with multiple soil series which were translated to a single parent material code corresponding to the dominant series, there is likely to be greater heterogeneity than indicated on the reference parent material map. Thus, while the derived parent material maps are used as the reference maps in this case, and are assumed to be correct, it must be recognised that they do not represent the absolute truth.

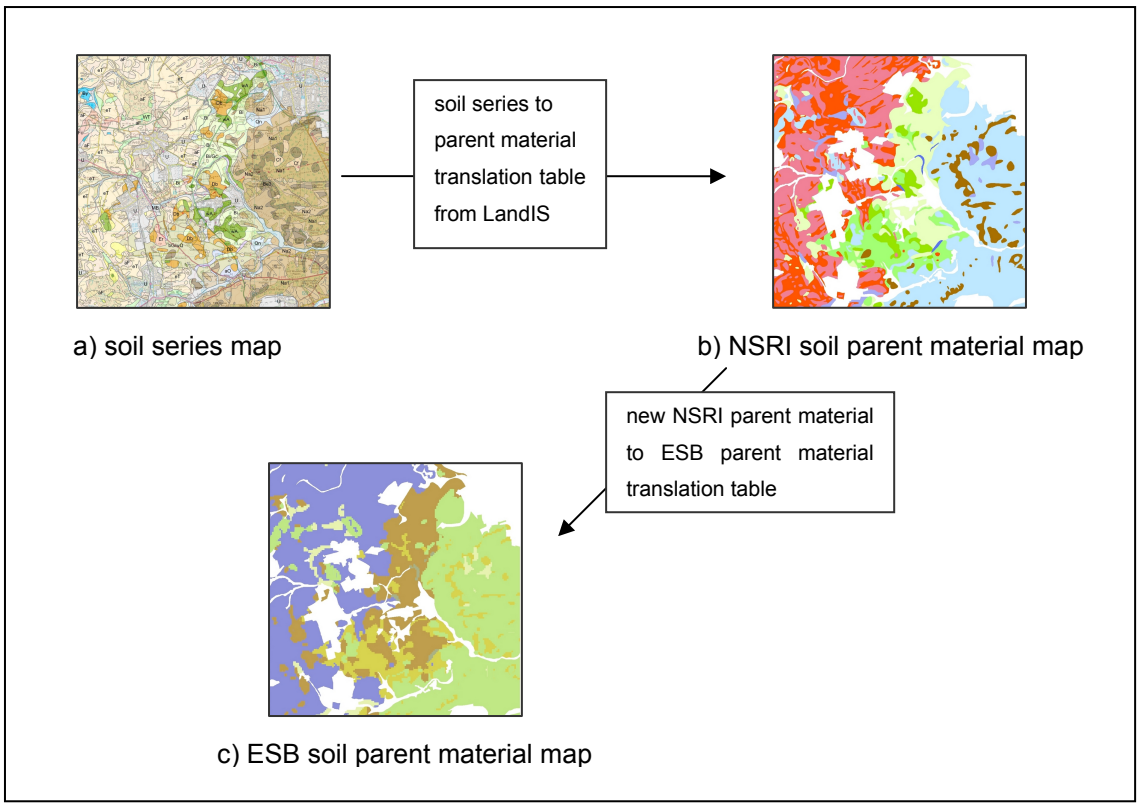


Figure 13 - Creation of the reference soil parent material maps

4.1.2 Geological maps (GEOLOGY)

Two geological layers were used in this research; a surface geology layer and a bedrock geology layer. Both layers are lithostratigraphic and were derived from the BGS 1:50,000 DiGMap digital datasets (British Geological Survey, 2007). In the coarse of this research, the surface geology was calculated from both the bedrock and superficial geology shapefile layers. If a superficial deposit was identified at a particular location, this was used as the surface geology. If no superficial geology was present, bedrock geology was used as the surface geology input. The bedrock geology input used only the bedrock component of DiGMap.

For the Worksop area, four geological sheets were combined to form a continuous layer across the study area. Edge matching problems occurred in this study area where insufficient rationalisation of the classifications had been performed by BGS, particularly with the calcareous mudstone unit (EDT_CAMD). The Yeovil and Needwood Forest sheets had no edge matching issues as they were fully contained within single geological sheets.

There is considerable geological diversity in all three of the study areas. Both Worksop and Needwood Forest have 15 geological units, while the Yeovil area has 26 (Table 4). Because of the lithostratigraphic mapping techniques employed for the geological mapping, the issues of map unit heterogeneity will not be as severe as for the soil maps (Sturdy, 1971) for the bedrock units. However, the superficial deposits have significantly greater lithological and textural variation across the mapped extent than the bedrock units, and are also more laterally discontinuous. As a result it is likely that superficial deposits will also have issues with map unit heterogeneity.

Concerns have also been expressed about the inconsistent mapping of superficial deposits in DiGMap (Palmer et al., 2007). The 1:625,000 scale DiGMap has consistent mapping of superficial deposits across the country, therefore, the use of this dataset was examined for suitability as an input into predictive parent material models. Unfortunately, serious geographic displacement issues of water bodies up to 1km were

found which could not be easily overcome. With this type of displacement, use of this dataset at scales of 1:50,000 was inappropriate, and so this layer was not used.

4.1.3 The National Soil Map (SOIL)

The 1:250,000 scale National Soil Map (NSRI, 2008a) is the most detailed soil dataset with national coverage across England and Wales. Due to the small scale of this map, the map units are soil associations, which are groupings of soil series found in association with each other in the landscape. There are multiple series within each association, and usually multiple soil parent materials within each association. Because of this, a translation of the National Soil Map to a parent material map would share the linework of the soil maps and contain numerous parent materials in each mapping unit.

The National Soil Map was mapped prior to the detailed soil mapping of the Yeovil area, but following the detailed mapping of the Needwood Forest and Yeovil areas. Therefore, the linework of the National Map is derived in part from the detailed mapping in these two areas, but is independent of the detailed mapping in the Yeovil area.

4.1.4 Slope maps (SLOPE)

Slope layers were derived from 5 m resolution Nextmap digital terrain models (DTM) (Intermap Technologies, 2002). The initial DTM processing steps are outlined in Appendix 7. The quantitative slope input was then classified according to the NSRI slope classification (Table 5) to allow input into the categorical probability model (see section 4.3)

Table 5 – Description of slope distribution in the three study areas

Note: This table is based on the soil survey handbook (Hodgson, 1997), also showing the percentage distribution of slope classes in the three study areas.

Description	Slope range (°)	Worksop	Needwood	Yeovil
level	0 - 1	6%	9%	10%
gentle	2 - 3	36%	27%	27%
moderate	4 - 7	38%	40%	40%
strong	8 - 11	10%	13%	13%
moderately steep	12 - 15	4%	5%	5%
steep	16 - 25	4%	4%	4%
very steep	26 - 36	1%	2%	2%
precipitous	36 - 90	0%	1%	1%

Analysis of the elevation and DTM derivatives revealed a number of errors, mainly artefacts from data processing by Intermap Technologies in the removal of surface features. These errors were typically found surrounding forest stands, as the sudden jump from the surface to the canopy creates a ‘precipitous’ slope. Issues such as these have been identified previously (Farmer, 2008). These errors, though limited in extent, are likely to affect the prediction of the soil parent material in these areas

Such errors could be corrected in an expert system model or by correcting the unprocessed DTM. Digital surface model (DSM) to DTM conversion is an area of image processing which requires extensive work and as such was beyond the scope of this project. As most of the identified errors were classified as precipitous slopes, the 'map purity' tables (Table 6) in the probability model, which describe the reliability of each evidence layer and class, were used to describe and account for the unreliable nature of the precipitous slope class.

4.2 Combining data layers and probabilities

The inconsistent scales of the vector datasets (reference maps, GEOLOGY, SOIL) contributed towards scale effects, such as slivers around the edges of polygons. These effects need to be borne in mind when analysing the data.

For the expert knowledge, data mining and combined methodologies which use more than one data input, it was necessary to use a model which allowed the integration of multiple evidence layers, (GEOLOGY, SLOPE, SOIL). The use of a probability model can also facilitate the incorporation of uncertainty in model outputs. This can allow more informed use of the resulting maps by end users. Because of the categorical nature of the mapped evidence layers, it was decided that a probability model based upon Bayes theorem would provide the necessary features to combine a range of evidence layers.

The Expector software package (Corner et al., 2002) was identified as a potentially suitable model, as it allows integration of both qualitative expert knowledge and quantitative data. Additionally, this approach allowed full visibility of the input data and calculations at all stages. However, initial attempts to use the software and associated models raised a number of concerns. Foremost amongst these were issues with the algebraic derivation of the equations used in Expector. As the equations stood, only the relative likelihood of each hypothesis was provided, rather than the desired actual

probabilities. The revised model is explained briefly below and in more detail in Farewell and Farewell (2010) which is presented in Appendix 5.

4.3 Probability model

The probability model created to calculate the probability of a pixel belonging to a range of parent material classes was derived from Bayes' Theorem.

$$P(H_k \mid E_1,..,E_n) = \frac{P(E_1,...,E_n \mid H_k)P(H_k)}{P(E_1,...,E_n)} \qquad [\ 1\ ]$$

Where H = Hypothesis (e.g. parent material class) and E = Evidence (e.g. GEOLOGY, SLOPE, SOIL).

Equation [1] was rewritten as Equation [2], adding the possibilities of weighting (W) the layers and assuming the following conditions:

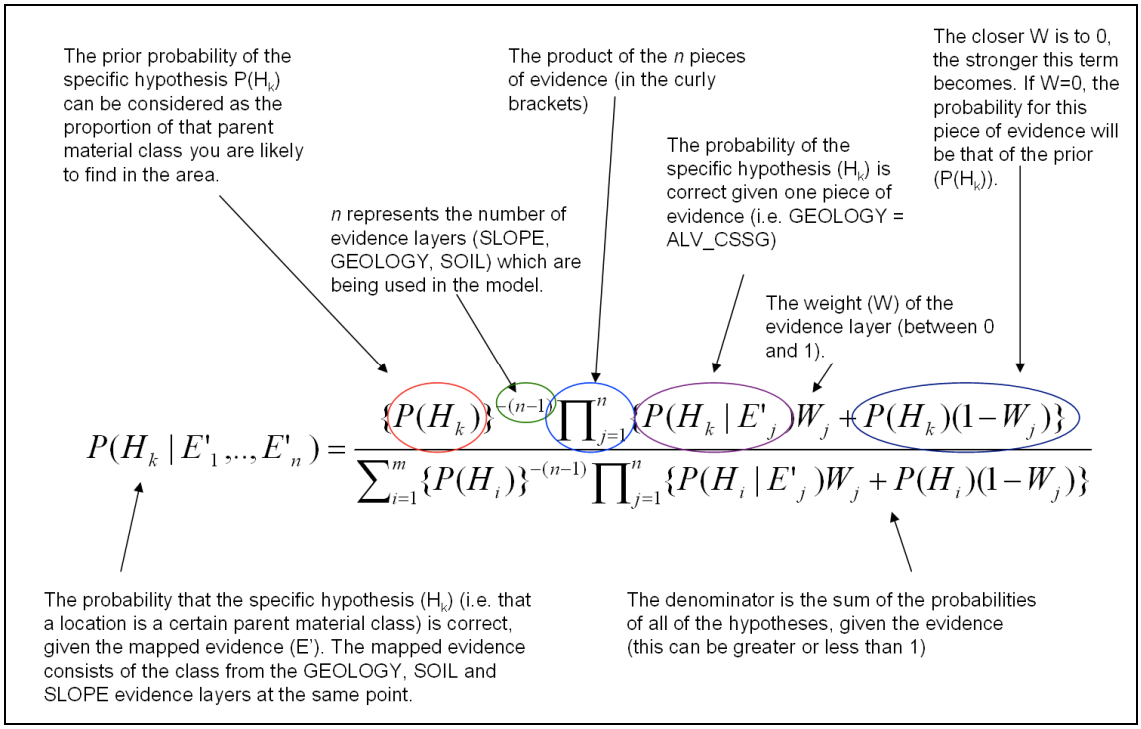
1. The *m* hypotheses are exhaustive and mutually exclusive (Law of Total Probability)
2. The individual pieces of evidence are independent, conditional on any given hypothesis H_i.

$$P(H_k \mid E'_1,..,E'_n) = \frac{\{P(H_k)\}^{-(n-1)} \prod_{j=1}^{n} \{P(H_k \mid E'_j)W_j + P(H_k)(1-W_j)\}}{\sum_{i=1}^{m} \{P(H_i)\}^{-(n-1)} \prod_{j=1}^{n} \{P(H_i \mid E'_j)W_j + P(H_i)(1-W_j)\}} \qquad [\ 2\ ]$$

Equation [2], which is a corrected and expanded version of the Expector Model (Corner et al., 2002; Farewell and Farewell, 2010) writes the desired probability in terms of the probabilities $P(H_i|E_j)$ of the various hypotheses, given the individual pieces of evidence, and the overall probabilities $P(H_i)$ of the hypotheses. This equation calculates the probability that the particular hypothesis (H_k) is true, given *n* pieces of evidence (E_n) for each hypothesis (i.e. GEOLOGY = MMG-MDST (a mudstone unit), SLOPE = moderate). The equation is described in Figure 14.

The number of hypotheses is equal to the number of parent material classes being predicted. For example, Hypothesis 1 (H_1) may be: "This location has a parent material of EiE1", Hypothesis 2 (H_2): "This location has a parent material of BhB1", etc…

Figure 14 – Explanation of the probability model



4.3.1 The probability model inputs

A VBA (visual basic for applications) macro in Microsoft Excel was created to run the probability models. An example is provided in Digital Appendix 2. The following inputs were required:

1) Prior probabilities of the hypotheses P(H)

These are, in effect, the % extent of each parent material class within the study area.

Then for each evidence layer (e.g. GEOLOGY, SLOPE, SOIL), the following inputs are required.

2) Prior probabilities of the evidence layers P(E)

These are, in effect, the % extent of the evidence layer class in the study area.

3) Evidence layer map purity table P(E|E')

This table provides a measure of confidence for a given evidence layer. It considers how trustworthy the map is; how closely it matches what is found in the field (Table 6). If field sampling was undertaken to test the purity of the evidence layer, individual class pairs could be weighted on the basis of the sample. The main soil mapping units on another 1:25,000 scale soil map were analysed for homogeneity of soil profiles within the drawn polygons (Sturdy, 1971). Correct identification ranged between 47 and 100% with the majority in the range of 60 to 75%.

No field sample was undertaken as part of this research and so a confidence level of 0.95 was applied for each class in the GEOLOGY and SOIL inputs, with the remaining 0.05 split between the other classes, on the assumption that the map would be correct 95% of the time. This higher level was chosen so as to minimise the influence of the prior probabilities on the model outputs as the percentage distribution among commonly misclassified units was unknown.

The SLOPE input was found to contain errors, particularly in the identification of precipitous classes. Thus, for this input, the "precipitous" class was given less confidence. Additionally, for other slope classes, because of the more continuous nature of the slope categories (Table 5), each class pair (e.g. "Gentle, Gentle") was rated with 0.8 and the remaining proportion divided between the flanking classes (e.g. "Level, Gentle" and "Moderate, Gentle").

Table 6 - An example map purity (P(E|E) table

Note: GEOLOGY map purity table from Worksop area. Because no knowledge was available of the trustworthiness of the geology map in this example, an assumed confidence of 0.95 was applied for each class.

Field Class \ Map Class	ALV-CSSG	BTH-DOLM	CDF-CAMD	CDF-DOLO	EDT-CAMD	EDT-MDSD	EDT-SDST	GFDMP-SAGR	HEAD-CSSG	LNS-SDST	NTC-PEST	RTD1-SAGR	TILMP-DMTN	WBY-MDST
ALV-CSSG	0.950	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
BTH-DOLM	0.004	0.950	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
CDF-CAMD	0.004	0.004	0.950	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
CDF-DOLO	0.004	0.004	0.004	0.950	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
EDT-CAMD	0.004	0.004	0.004	0.004	0.950	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
EDT-MDSD	0.004	0.004	0.004	0.004	0.004	0.950	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
EDT-SDST	0.004	0.004	0.004	0.004	0.004	0.004	0.950	0.004	0.004	0.004	0.004	0.004	0.004	0.004
GFDMP-SAGR	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.950	0.004	0.004	0.004	0.004	0.004	0.004
HEAD-CSSG	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.950	0.004	0.004	0.004	0.004	0.004
LNS-SDST	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.950	0.004	0.004	0.004	0.004
NTC-PEST	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.950	0.004	0.004	0.004
RTD1-SAGR	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.950	0.004	0.004
TILMP-DMTN	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.950	0.004
WBY-MDST	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.950

4) The joint probability of hypothesis and evidence P(H,E)

This table represents the probability of, both the hypothesis and evidence being found at the same point. For example, both GEOLOGY = HEAD-CSSG and Hypothesis = BhB1 being present at the same point; e.g. P(BhB1, HEAD-CSSG) = 0.01).

Table 7 – An example joint probability (P(H,E)) table

Note: This table shows the likelihood of an evidence class and hypothesis class being found at the same point. For clarity, cells with a value of 0 have been removed. The 0.00 values in this table represent low probabilities, less than 0.01. The soil parent material classes, e.g. BhB1, are described in **Appendix 2**.

		Hypotheses								
		BhB1	**BhB2**	**BoB2**	**EaE1**	**EfE1**	**EiE1**	**FiF1**	**FiF2**	**FqF1**
		0.26	*0.22*	*0.10*	*0.00*	*0.04*	*0.20*	*0.13*	*0.03*	*0.02*	*P(H)*
Evidence	**ALV-CSSG**	*0.01*	0.00	0.00		0.00	0.00	0.00	0.00		
	BTH-DOLM	*0.00*	0.00	0.00					0.00	0.00	
	CDF-CAMD	*0.01*	0.00	0.00			0.00	0.00	0.00		
	CDF-DOLO	*0.51*	0.25	0.20	0.00		0.02	0.01	0.02	0.00	0.00
	EDT-CAMD	*0.00*	0.00	0.00					0.00		0.00
	EDT-MDSD	*0.13*	0.00	0.00	0.01	0.00	0.00	0.01	0.07	0.02	0.01
	EDT-SDST	*0.05*			0.00			0.02	0.02	0.00	
	GFDMP-SAGR	*0.01*	0.00	0.00	0.00		0.00	0.00	0.00		0.00
	HEAD-CSSG	*0.03*	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
	LNS-SDST	*0.07*		0.00	0.01			0.05	0.00	0.00	0.00
	NTC-PEST	*0.19*			0.07		0.01	0.10			
	RTD1-SAGR	*0.00*					0.00				
	TILMP-DMTN	*0.01*	0.00	0.00				0.00	0.00	0.00	0.00
	WBY-MDST	*0.00*							0.00		
		P(E)									

P(H,E) is in fact unknown when populating such tables from expert knowledge and therefore needs to be calculated from P(E|H) (see Table 8), which is the conditional probability of the evidence, given the hypothesis.

Table 8 – An example conditional probability (P(E|H)) table

Note: This table shows the likelihood of an evidence class being found, assuming the hypothesis class is known. This is the format in which expert knowledge can be entered into the probability model. For clarity, cells with a value of 0 have been removed. The 0.00 values in this table represent low probabilities, less that 0.01. The soil parent material classes, e.g. BhB1, are described in **Appendix 2**.

			Hypotheses								
			BhB1	**BhB2**	**BoB2**	**EaE1**	**EfE1**	**EiE1**	**FiF1**	**FiF2**	**FqF1**	
			0.26	*0.22*	*0.10*	*0.00*	*0.04*	*0.20*	*0.13*	*0.03*	*0.02*	*P(H)*
Evidence	**ALV-CSSG**	*0.01*	0.00	0.01		0.14	0.07	0.01	0.02			
	BTH-DOLM	*0.00*	0.00	0.00						0.05	0.02	
	CDF-CAMD	*0.01*	0.00	0.01			0.06	0.00	0.01			
	CDF-DOLO	*0.51*	0.95	0.92	0.01		0.53	0.04	0.18	0.14	0.21	
	EDT-CAMD	*0.00*	0.00	0.00					0.01		0.00	
	EDT-MDSD	*0.13*	0.01	0.01	0.07	0.66	0.00	0.05	0.55	0.74	0.60	
	EDT-SDST	*0.05*			0.03			0.12	0.15	0.01		
	GFDMP-SAGR	*0.01*	0.00	0.00	0.02		0.01	0.01	0.00		0.00	
	HEAD-CSSG	*0.03*	0.02	0.03	0.00	0.20	0.04	0.01	0.06	0.02	0.00	
	LNS-SDST	*0.07*		0.00	0.14			0.24	0.01	0.00	0.17	
	NTC-PEST	*0.19*			0.72		0.28	0.51				
	RTD1-SAGR	*0.00*					0.01					
	TILMP-DMTN	*0.01*	0.01	0.02				0.00	0.01	0.05	0.00	
	WBY-MDST	*0.00*							0.00			
		P(E)										

5) Evidence layer weights (optional)

The ability to weight the evidence layers (from 0 to 1) was added to the probability model; see Equation [2]. The default value for the weight is 1. A weight of 0 will effectively remove a layer from the model. This weighting provides a simple mechanism by which models could be altered to derive more or less emphasis from a particular evidence layer. Attempts have also been made to use this weighting as a simple alternative to the map purity table (Table 6), in the absence of detailed knowledge regarding map purity (Farewell and Farewell, 2010). Initial results are encouraging, but this requires additional investigation beyond this research.

4.3.2 Model outputs

From these inputs, a number of calculations are performed, and $P(H_k|E_1,..,E_n)$ (the probability of each hypothesis being found, given the evidence layers) is output for each hypothesis given the evidence layers. These probabilities are then compared and a most likely hypothesis determined for each combination of evidence classes (Table 9). The most likely hypothesis is presented in the mapped results (for example, Figure 19) and used in the derivation of summary statistics and metrics, which will be discussed in the next chapter.

Table 9 - An example results table ($P(H|E_1,E_2,E_3)$)

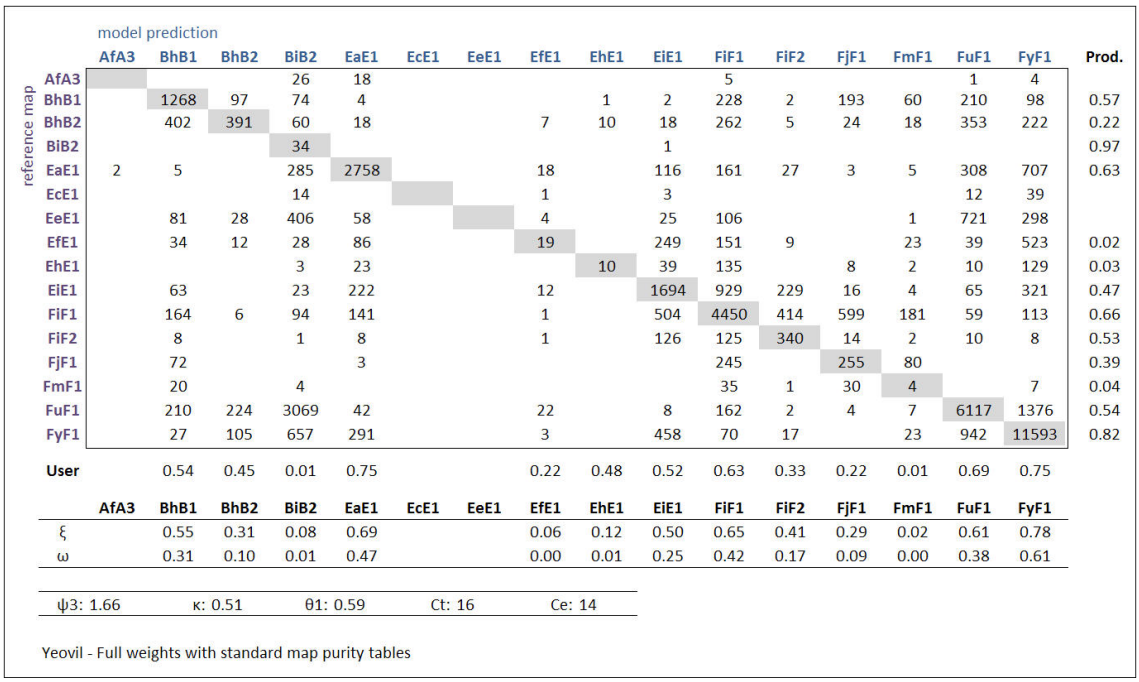
Note: only the first 13 rows (of 1232 rows are presented). The probabilities of each parent material class are calculated (blue headings) based on the combination of evidence layer classes (red headings). The most likely hypothesis is highlighted in grey (note: the blue highlighting indicates that the difference is not always large) and summary information, along with a join field for linking to the GIS shapefiles is provided under the green headings. The soil parent material classes, e.g. BhB1, are described in **Appendix 2**.

evidence layer fields			probabilities of the hypotheses - P(H|E1,E2,E3)									summary output		
GEOLOGY	SLOPE	SOIL	BhB1	BhB2	BoB2	EaE1	EfE1	EiE1	FiF1	FiF2	FqF1	Most Likely	Probability	Join Field
ALV-CSSG	level	51101	0.05	0.50	0.00	0.01	0.32	0.03	0.09	0.00	0.00	BhB2	0.50	ALV-CSSGlevel51101
ALV-CSSG	level	54102	0.01	0.16	0.00	0.20	0.04	0.13	0.36	0.07	0.03	FiF1	0.36	ALV-CSSGlevel54102
ALV-CSSG	level	54118	0.00	0.04	0.00	0.01	0.04	0.74	0.15	0.01	0.01	EiE1	0.74	ALV-CSSGlevel54118
ALV-CSSG	level	54300	0.00	0.00	0.00	0.34	0.01	0.22	0.42	0.00	0.00	FiF1	0.42	ALV-CSSGlevel54300
ALV-CSSG	level	55101	0.02	0.33	0.03	0.01	0.33	0.13	0.15	0.00	0.00	BhB2	0.33	ALV-CSSGlevel55101
ALV-CSSG	level	55102	0.00	0.01	0.01	0.01	0.31	0.63	0.04	0.00	0.00	EiE1	0.63	ALV-CSSGlevel55102
ALV-CSSG	level	57206	0.00	0.03	0.00	0.00	0.96	0.00	0.01	0.00	0.00	EfE1	0.96	ALV-CSSGlevel57206
ALV-CSSG	level	57212	0.00	0.16	0.00	0.01	0.03	0.58	0.22	0.00	0.00	EiE1	0.58	ALV-CSSGlevel57212
ALV-CSSG	level	71103	0.00	0.02	0.00	0.25	0.01	0.01	0.70	0.00	0.00	FiF1	0.70	ALV-CSSGlevel71103
ALV-CSSG	level	71301	0.00	0.04	0.00	0.00	0.01	0.04	0.91	0.00	0.00	FiF1	0.91	ALV-CSSGlevel71301
ALV-CSSG	level	82102	0.00	0.01	0.00	0.01	0.03	0.92	0.03	0.00	0.00	EiE1	0.92	ALV-CSSGlevel82102
ALV-CSSG	gentle	51101	0.04	0.48	0.00	0.01	0.34	0.03	0.10	0.00	0.00	BhB2	0.48	ALV-CSSGgentle51101
ALV-CSSG	gentle	54102	0.00	0.13	0.00	0.30	0.04	0.11	0.34	0.05	0.03	FiF1	0.34	ALV-CSSGgentle54102

4.4 Data analysis

The same analyses are undertaken for each methodology in this research. Firstly, to allow comparison between the modelled and reference parent material maps, a dense 60 m grid (a point shapefile) was attributed with parent material classes from both maps. A dense point grid was chosen for the analyses, as it allowed for easy integration of the point based SLOPE layer. The 60 m grid was chosen as this was the maximum

number of sample points that could be analysed in Microsoft Excel 2003 (65536 row limit). Comparisons were made with analyses on a 25 m grid and were found to be equivalent with those on a 60 m grid (Section 4.6). The records from the shapefile (one for each point on the grid) were loaded into Excel and a confusion matrix created (Figure 15) from which a range of map value and class value metrics are calculated. These metrics are now discussed.



model prediction

reference map	AfA3	BhB1	BhB2	BiB2	EaE1	EcE1	EeE1	EfE1	EhE1	EiE1	FiF1	FiF2	FjF1	FmF1	FuF1	FyF1	Prod.
AfA3				26	18						5				1	4	
BhB1		1268	97	74	4				1	2	228	2	193	60	210	98	0.57
BhB2		402	391	60	18			7	10	18	262	5	24	18	353	222	0.22
BiB2				34						1							0.97
EaE1	2	5		285	2758			18		116	161	27	3	5	308	707	0.63
EcE1				14				1		3					12	39	
EeE1		81	28	406	58			4		25	106			1	721	298	
EfE1		34	12	28	86			19		249	151	9		23	39	523	0.02
EhE1				3	23				10	39	135		8	2	10	129	0.03
EiE1		63		23	222			12		1694	929	229	16	4	65	321	0.47
FiF1		164	6	94	141			1		504	4450	414	599	181	59	113	0.66
FiF2		8		1	8			1		126	125	340	14	2	10	8	0.53
FjF1		72			3						245		255	80			0.39
FmF1		20		4							35	1	30	4		7	0.04
FuF1		210	224	3069	42			22		8	162	2	4	7	6117	1376	0.54
FyF1		27	105	657	291			3		458	70	17		23	942	11593	0.82
User		0.54	0.45	0.01	0.75			0.22	0.48	0.52	0.63	0.33	0.22	0.01	0.69	0.75	

	AfA3	BhB1	BhB2	BiB2	EaE1	EcE1	EeE1	EfE1	EhE1	EiE1	FiF1	FiF2	FjF1	FmF1	FuF1	FyF1
ξ		0.55	0.31	0.08	0.69			0.06	0.12	0.50	0.65	0.41	0.29	0.02	0.61	0.78
ω		0.31	0.10	0.01	0.47			0.00	0.01	0.25	0.42	0.17	0.09	0.00	0.38	0.61

ψ3: 1.66 κ: 0.51 θ1: 0.59 Ct: 16 Ce: 14

Yeovil - Full weights with standard map purity tables

Figure 15 – An example confusion matrix and associated analyses
Note: The columns represent the model prediction, for example, EcE1 and EeE1 were not predicted by this model. The rows to the left of the confusion matrix represent the 'true' soil parent material, according to the reference map. The soil parent material classes, e.g. AfA3, are described in **Appendix 2**.

4.5 Qualities of valuable parent material maps

Different maps serve different purposes and therefore, there is a need to understand the purpose for which a map is created. The desired attributes of a parent material map useful for input into environmental models at approximately 1:50,000 scale are explicitly stated for this research, helping define the criteria by which success of the resulting parent material map is assessed.

In the context of this research, a useful parent material map would have:

- Numerous parent material classes, which are representative of the area
- Clearly defined, highly specific parent material classes, related to soil types
- Parent material classes which accurately represent the geographic reality
- A high overall accuracy.

Ideally, the modelled parent material map would perfectly reflect the spatial distribution of a large number of clearly defined classes of parent material, across all scales. In reality, this is not feasible due to limits on time, costs and the validity of the input datasets, so compromises are inevitable. One example may be the widening of a class definition to reduce misclassification and increase overall map accuracy.

Consistent and quantifiably supportable conclusions and recommendations are sought regarding the best methods for creating a useful soil parent material map. In this regard, a number of metrics, statistics and analyses to allow quantitative comparisons to be effectively made between different methodologies and model runs were required

At this stage, only qualitative indicators of map value have been stated. Now, using these qualitative statements, the quantitative assessment of the value of an individual parent material *class* will be discussed, after which the measurement of the overall success or value of the *map* will be discussed.

4.5.1 Individual parent material class value analyses

From initial spatial analysis of model tests using confusion matrices (Rosenfield and Fitzpatrick-Lins, 1986; Landis and Koch, 1977), it was found that certain soil parent material classes provided accurate predictions, while others were consistently misclassified (see FuF1:BiB2 pairing, Figure 15). The ability to quantitatively state which parent material units were predicted well was desired so that confidence levels could be applied to any resulting map. This provides knowledge of the propagation of

errors in the resulting maps (Goodman, 1960) and allows an assessment of class or map value to be made.

For the purposes of this research, a valuable parent material map *unit* was defined as having the following characteristics.

- A highly specific class definition, with closely defined characteristics of the soil parent material.
- Accurate spatial representation of 'reality', where under and over prediction are minimised and where there is good spatial agreement with reference maps (Figure 16 b).

4.5.1.1 Producer and user accuracies

Producer (A_p) and user (A_u) accuracies are commonly used descriptive statistics for classes in remote sensing and related disciplines, and were calculated for each parent material unit, in each test. They were, however, found not to be entirely fit for the purposes of this assessment of 'class value' as each statistic only relates to a portion of the success of the prediction, as explained below and in Figure 16.

The user accuracy is calculated by dividing the area of agreement by the total area of that map unit shown on the model-derived map. If the user accuracy is low, this can indicate that there is over prediction of this parent material unit. If the user accuracy is high, but the producer accuracy is low, (Figure 16 (a)) this unit is under predicted and of little value.

The producer accuracy is calculated by dividing the area of unit agreement by the total area of that unit from the reference map. If the producer accuracy is low, this indicates that there is a lot of under prediction of this parent material unit. If the producer accuracy is high, but the user accuracy is low, (Figure 16 (d)), this unit is over predicted and also of little value.

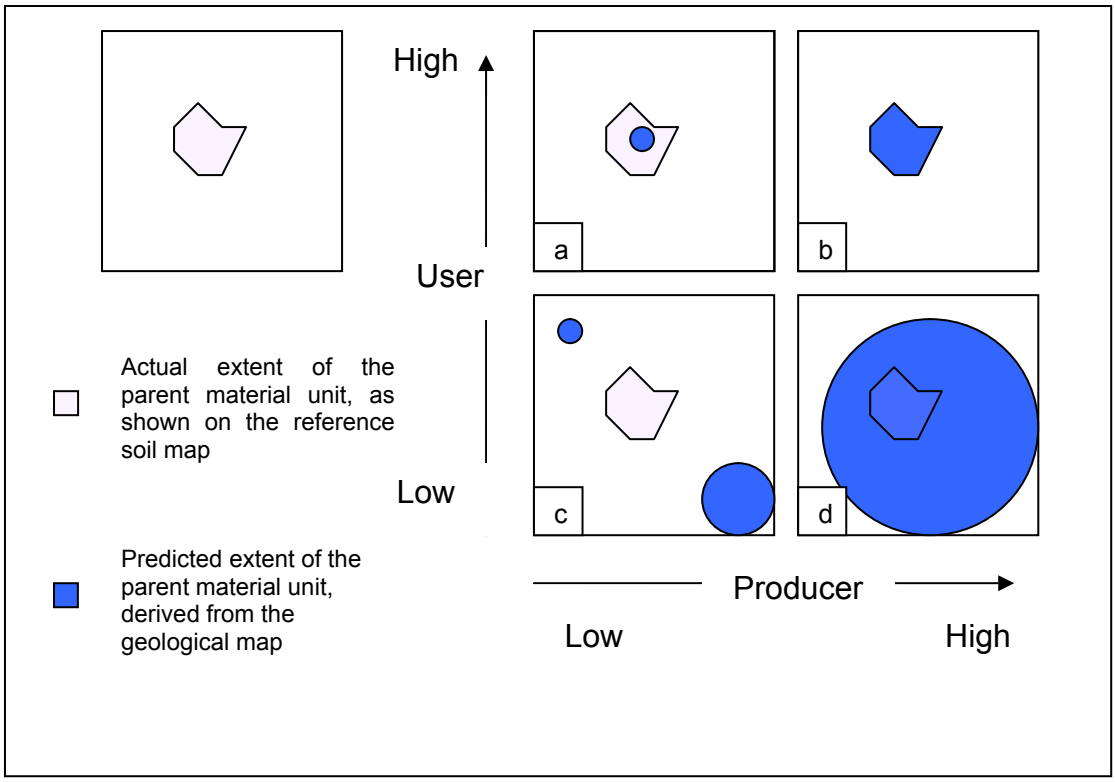


Figure 16 - Producer and user accuracies
a) under prediction of the map unit. $A_p \approx 0.2$, $A_u = 1.0$ b) perfect prediction of the map unit. $A_p = 1.0$, $A_u = 1.0$ c) misclassification of the map unit. $A_p = 0.0$, $A_u = 0.0$ d) over prediction of the map unit. $A_p = 1.0$, $A_u \approx 0.2$

The ability to indicate the relative value of different parent material classes for input into other environmental models is desired, as this allows a measure of uncertainty to be associated with the resulting map output. Both user and producer accuracies independently provide some information as to the success of the model identifying the soil parent material. It is possible, however to have a very high producer accuracy and a very low user accuracy, (and vice-versa; see Figure 2, (a) and (d)). Such units would be of little use in a parent material map. Therefore a new combined assessment of class value (ξ) using the geometric mean of the producer and user accuracies has been developed to provide information on the success of prediction of each parent material class.

4.5.1.2 Xi (ξ) simple class value indicator

Geometric means have been used in machine learning and artificial neural networks to aid more accurate prediction of minority classes (Kubat et al., 1998). The ξ class value indicator was developed in the course of this research to provide an indication of the overall value of a particular parent material class, regardless of its extent. ξ is calculated as the geometric mean of the user (A_u) and producer (A_p) accuracies for the parent material unit in question.

The ξ class value indicator is calculated as

$$\xi = \sqrt{A_u A_p} \qquad [3]$$

The ξ class value indicator can take a value between 0 (no value) and 1 (high value, perfect classification success). ξ can only be close to 1 if both A_u and A_p are close to 1. Because of the intuitive scale and ease of comparison between classes, the ξ class value indicator was found to be very useful in describing the relative successes of the parent material classes in this research, and has also been used extensively in related work (Palmer et al., 2007). Because of its inherent statement of the value of a class, ξ may also be used as a map unit weighting tool for digital mapping systems or for knowledge of error propagation.

4.5.1.3 Omega (ω) weighted class value indicator

In cases where there is a significant amount of confusion between parent material classes, it may be advantageous to amalgamate or combine classes to form a broader parent material class. These broader classes, following amalgamation, often achieve a higher level of predictive success and thus, higher class value (ξ). However, when classes are combined, they become less useful as predictors of *specific* parent materials. Nevertheless, a balance must be achieved between predictive success and the degree of specificity of the class.

A simple solution considered to include a measure of class size to this indicator was to divide ξ by the number of members of the class in question (c). However, in the context of this research, this was found to place too much emphasis on the value of classes with only one unit. Therefore, a number of different metrics of this type, with slightly different emphases, were tested in order to discover which metric produced results which most closely matched what was intuitively felt to be a 'valuable' class. The selected metric is presented in Equation [4] and its responses for (higher value) map units with one class up to (lower value) broad map units with five units are graphically shown in Figure 17.

$$\omega = \frac{\xi_i^{\,2}}{\sqrt{c_i}} \qquad [\,4\,]$$

Where *c* is the number of parent materials in an amalgamated class

The chosen emphases within the metric in Equation [4] were subjective. For example, ξ is squared to penalise low class values, and hence, reward classes with high values. Such metrics can be modified for situations where certain issues may be more important than others. For example, some consideration of the similarity of classes may be important in some situations, in which case *c* could be based on a measure of taxonomic distance (Minasny and McBratney, 2007). In the context of this research, however, the metric appeared to be the most appropriate. ω for each map unit also forms an integral part of the overall map value metric (ψ_3).

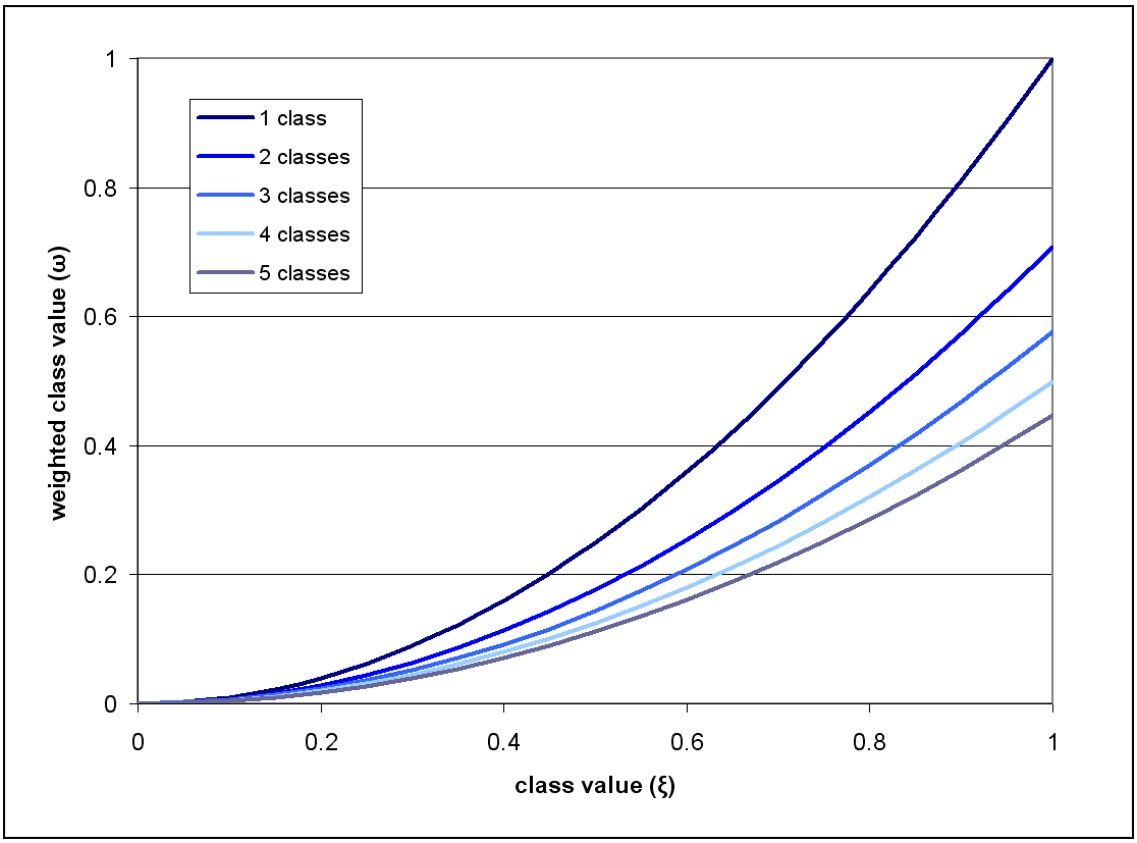


Figure 17 - the relationship between class value (ξ) and weighted class value (ω) for map units with 1,2,3,4 and 5 component classes.

4.5.2 Whole map value analyses

In order to determine the success of the models across the whole of each study area, two traditional accuracy assessments were initially used, the overall accuracy (θ_1) and Fleiss's kappa statistic (κ) (Fleiss, 1971; Landis and Koch, 1977; Hudson and Ramm, 1987).

The overall accuracy (θ_1) of the study area is the sum of the areas of agreement between the parent material maps derived from the soil and geological layers, divided by the total area of the map. This assessment includes both random and non-random areas of agreement, and has a range of 0 (no agreement) to 1 (total agreement).

The Fleiss variant of Cohen's kappa (κ) statistic is a statistical measure of inter-rater (or inter-model) reliability. Like θ_1, κ has a range of 0 to 1 but it is more sophisticated. It is used to provide a measure of only the non-random agreement between modelled results and the 'truth'. It can be considered as being the overall agreement (θ_1) minus the chance agreement. In this research, κ was initially used to provide an indication of the overall model success with reference to the reference soil map. For the method of calculating kappa, see Hudson and Ramm (1987).

Delta kappa (Hudson and Ramm, 1987) is calculated using the kappa statistic and variance of kappa from two tests, and has been used to test for significant difference between tests. While delta kappa was considered for use in this research it was rejected as such large sample sizes were used that the variance of kappa was always very close to zero. This made every test significantly different from every other, and rendered this test unsuitable for the purposes required by this research. It is vital to separate statistical significance from scientific importance.

4.5.3 Issues with kappa (κ) and the overall accuracy (θ_1)

During the research, it was determined that neither the overall accuracy (θ_1) nor the kappa statistic (κ) were providing a reliable assessment of the overall model success. Neither assessment provided the full picture as to the many different factors which contribute to a 'useful' parent material map, and by seeking to maximise the value of one assessment, such as κ, could often lead to a less useful map, for example, a map with only one or two very broad map units.

An extreme example of this issue would be the amalgamation of all classes to achieve a θ_1 value of 1. Such a map would be unsuitable as a parent material map as there would be no differentiation between parent material classes. While examples are not so extreme using κ, as this assessment considers the total number of classes, similar issues

do arise. Therefore a new metric of map value (ψ_3) was derived which considered a number of different factors contributing to a valuable map.

4.5.4 The map value psi (ψ_3) metric

The desired attributes of a useful parent material map which have been stated (see 4.5) were used to create a new quantitative metric for assessing the usefulness of the resulting parent material map. Based on the stated attributes, the new map value metric (ψ_3) was calculated as shown in Equation [5]. It is annotated for clarity in Figure 18.

$$\Psi_3 = \sum_{i=1}^{n} \left(\frac{\xi_i^{\,2}}{\sqrt{c_i}} \right) \theta_1 \qquad [\ 5\ ]$$

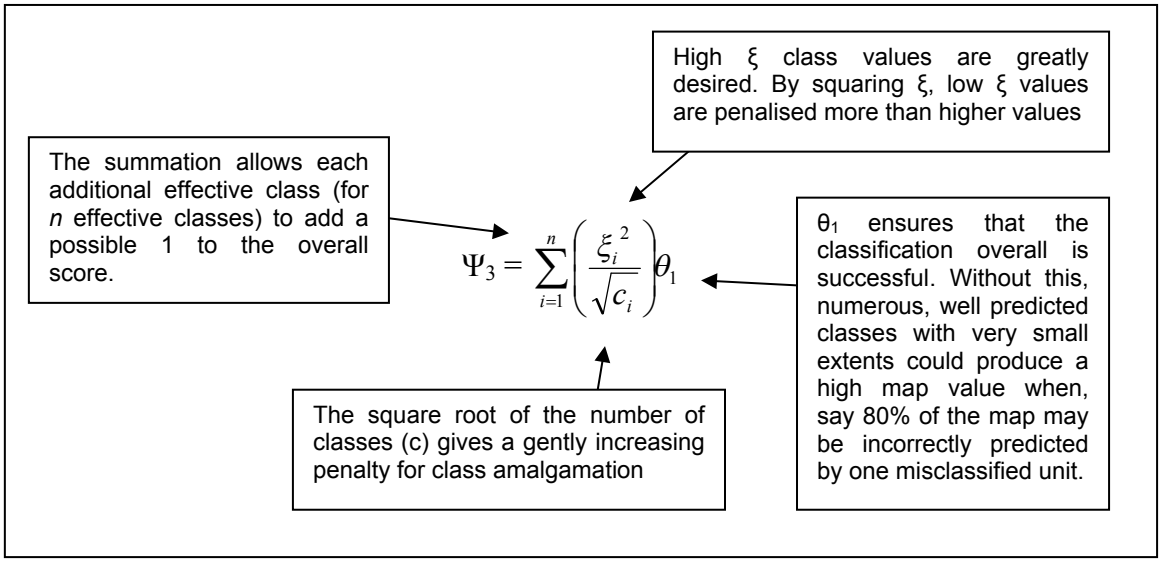


Figure 18 – Description of the ψ_3 map value metric

The map value metric (ψ_3) incorporates a number of the different measures of map 'usefulness', providing a single-number comparative metric which is suitable for comparisons between tests of all methodologies, within the same study area. The map value metric is not suitable for comparisons between study areas, as different

geographic areas will have different levels of geo-diversity leading to a varying number of potential soil parent materials within the study areas.

As with the integrated assessment of class value (ω), a number of different weights and emphases of the ψ_3 metric were tested in order to create the indicator which most closely matched what was intuitively felt by the author to be a 'valuable' parent material map. Again, the emphases were subjective and different research may demand different emphases. In the context of this research, however, the ψ_3 metric as presented in Equation [5] appeared to be the most appropriate single value comparative metric. Nevertheless, it is always informative to consider ψ_3 with reference to the other metrics discussed, so a full understanding can be obtained of how the single comparative ψ_3 value was reached.

The ψ_3 metric can be used to compare map value between different parent material classifications. For maps which use broader classifications, for example the ESB or simplified NSRI classifications (Appendix 2), as part of the ψ_3 calculations, the number of fully detailed classes (c) which make up the broader class are considered. This allows rational comparison of success between the methodologies, even between classifications with different levels of detail.

The maximum map value achievable is dependant on the number of parent material classes in the study area. Worksop has 9, Needwood Forest has 11 and Yeovil has 17. Because of this variable maximum value, the map value metrics are not comparable between maps of different regions or study areas. However, the alternative measures of overall accuracy (θ_1) and the kappa statistic (κ) can provide comparisons of aspects of value between study areas.

In the context of this research, a 'more valuable map' has a higher map value (ψ_3) than a 'less valuable map'.

4.5.5 The derivation and application of the ψ_3 metric

The map value (ψ_3) metric was one of multiple map value metrics investigated for this research. Initially, a simplistic summation of the un-weighted class values was tested, as in Equation [6].

$$\Psi_0 = \sum_{i=1}^{n} \zeta_i \qquad [\ 6\ ]$$

This equation was found to encourage amalgamation of classes to the detriment of class detail. Furthermore, no consideration was given to the overall accuracy of the map. For example, if a map had a low overall accuracy (say 0.15) but a number of well-predicted classes with very limited extents, this map would achieve a higher ψ_0 value than a map with a higher overall accuracy (say 0.50) but with the classes less well predicted. These two concerns were addressed in the ψ metric in Equation [7].

$$\Psi = \sum_{i=1}^{n} \left(\frac{\xi_i}{\sqrt{c_i}} \right) \theta_1 \qquad [\ 7\ ]$$

Equations [8], [9] and [10] applied different weightings to the various components of the map value equations.

$$\Psi_1 = \sum_{i=1}^{n} \left(\frac{\xi_i}{\sqrt{c_i}} \right) \theta_1^{\,2} \qquad [\ 8\ ]$$

$$\Psi_2 = \sum_{i=1}^{n} \left(\frac{\xi_i^{\,2}}{\sqrt{c_i}} \right) \theta_1^{\,2} \qquad [\ 9\ ]$$

$$\Psi_3 = \sum_{i=1}^{n} \left(\frac{\xi_i^{\,2}}{\sqrt{c_i}} \right) \theta_1 \qquad [\ 10\ ]$$

The metrics in Equations [8], [9] and [10] were calculated for all tests in this research, but it was the ψ_3 metric which most commonly closely matched what was

intuitively felt to be the most valuable maps in this research. Therefore it was this map value metric which is used as the primary measure of success throughout this research.

Because the ψ_3 metric incorporates the numerous measures of the aspects of a valuable map discussed above, the relationship of this metric to each individual component of the equation is influenced by the other components. Therefore, while the ψ_3 metric has been found to be a useful indicator of map value in the context of this research, it is advised that the supplementary statistics of class value (ξ), the number of effective classes (C_e) and the overall accuracy are considered alongside this new metric, as this allows for a more detailed assement of the value of the map to be made

4.5.6 Effective classes (C_e) and total classes (C_t)

In this research, effective parent material classes (C_e) have been defined as those present in both the modelled parent material and the reference parent material map. It is possible that a parent material class may only be identified by either the reference or the modelled map. A unique list of all parent material classes from both reference and modelled maps defines the total number of classes (C_t).

4.6 Sample density for test analyses

Tests were performed comparing the difference in model results with a 25 m grid and a 60 m grid (Table 10). The 60 m grid was not resampled from the 25 m grid but was a separate grid. The Yeovil study area was used as this had the greatest diversity in soil, geology and parent material units.

Both sample grid tests produced the same κ and θ_1 values. This demonstrates that a 60 m spacing between sample points is as statistically robust as using a 25 m spacing for the purpose of model testing in this research. A couple of the parent material units, for

example, EfE1, correctly identified the parent material in the 25 m test where no correct pixels were found in the 60 m test. However, there was also an increase in the number of incorrectly classified pixels, so overall there was no change in the results.

Table 10 – Model test point density comparison

Note: κ = overall agreement, minus chance agreement, θ_1 = overall accuracy, C_t = total number of parent material classes identified in either soil map or modelled map, C_e = effective classes (those occurring in both the soil map and the modelled map)

Test point spacing	κ	θ_1	C_t	C_e	Sample points
60 m	0.37	0.50	17	11	55,529
25 m	0.37	0.50	17	16	318,869

4.7 The presentation of results in this research

For consistency, results are presented in a standardised manner throughout this research. The presentation of mapped and tabular results is discussed here, along with explanatory notes.

4.7.1 Mapped results

The output file (Table 9) for each model run was joined to a systematic 60 m sampled point shapefile containing the attribute data of all the input evidence layers. Systematic grids can under represent the extent of small mapping units. While a number of other types of sampling strategies were considered, (e.g. random or stratified sampling), it was demonstrated that all units except one very small unit in the Yeovil study area were identified with 60 m sampling. Furthermore, a systematic grid provides a consistent sampling matrix allowing simpler display of the sampled data. This grid allowed the success of the model to be assessed both visually, and by statistical analysis of the

sample data. Standardised colours were used in visual comparisons of the parent material classes.

The results of the models are graphically presented for simple comparison with the reference parent material map, and assessment of the success and confidence of the model. Four maps (a to d) are presented in each case (see the example in Figure 19):

a) the reference parent material map (if amalgamated classes are used in the model, these classes are amalgamated on this map as well)

b) the map resulting from the model run, showing the most likely parent material unit, given the evidence (Table 9).

c) Model and map agreement. If the most likely parent material is the same as that on the reference map, this is coloured green, if not, red. More sophisticated, fuzzy map comparison techniques are available (Visser and de Nijs, 2006) but were not used in these maps, as an absolute yes / no answer was sought for clarity.

d) Model prediction confidence or probability of the most likely hypothesis (Table 9). (The results from the first (data dictionary) methodology, where there is no model confidence or probability do not include this map).

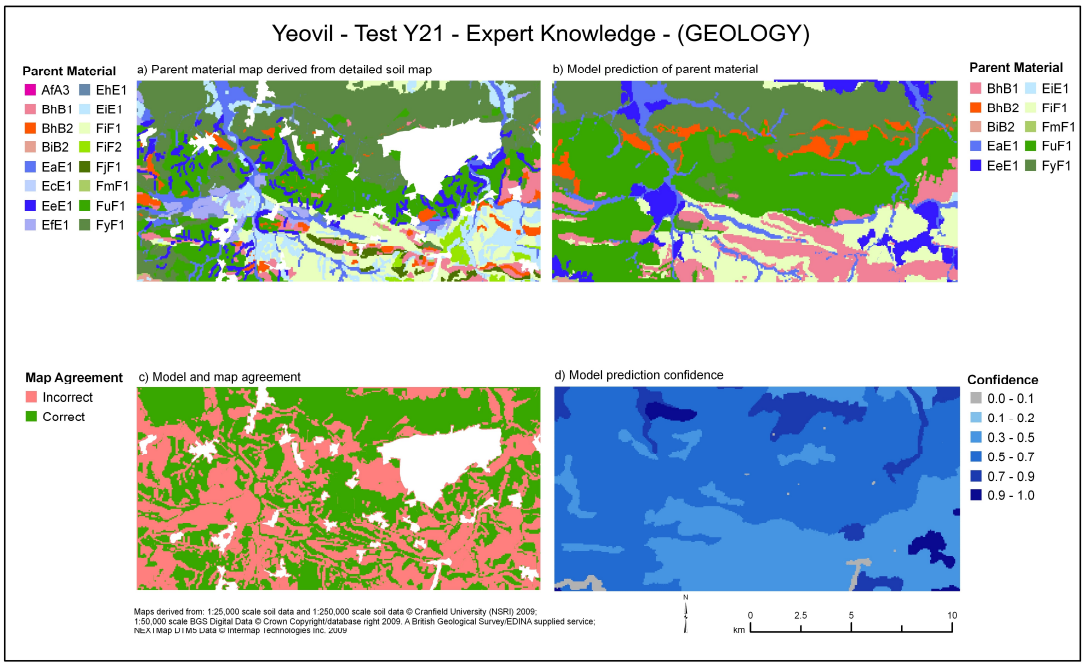


Figure 19 - Example presentation of mapped result
Note: the title displays the study area, test number, methodology and the evidence layers used.

To aid discussions, these maps are presented for quick reference at a reduced scale in the text. All mapped results are also presented for easy comparison in Appendix 4.

4.7.2 Result Tables

Many unique tests were run during the course of each methodology. Theses might vary in input data, classifications, or weightings. Key tests are presented in tabular format for each methodology, and each study area. As previously discussed, many statistics and metrics can be used to describe the relative success of different models and methodologies. In the course of this research and in the following discussions, the map value (ψ_3) metric is used as the primary descriptor of model success. Eleven supporting summary details, statistics and metrics are also presented alongside this value to describe the test in more detail. An example table is fully described in Table 11.

Table 11 – An example results table

Yeovil	Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
PM_LITH (surf.)	(A1)	0.11	**0.13**	**0.15**	19	3	3	3	3	-	-	2	83%	Y1
PM_LITH (surf.)	(A2)	0.11	**0.14**	**0.23**	17	5	4	4	4	1	-	3	72%	Y2
PM_LITH (surf.)	(A3)	0.62	**0.76**	**1.39**	11	6	6	6	5	1	4	6	1%	Y3

1. **Method** – Description of the test
2. **(κ)** – Fleiss's variant of Cohan's kappa statistic: the amount of agreement between the modelled map and the 'truth', minus the chance agreement.
3. **(θ_1)** – The overall accuracy of the map: the proportion of the map correctly predicted
4. **(ψ_3)** – The map value metric, described above
5. **Total Classes** –The total number of unique classes in both the model results and the 'true' parent material map.
6. **Effective Classes** – The number of effective classes, those classes which occur on both the modelled map and the 'true' map.
7. **C ξ > 0.2** – The number of classes which have a class value (ξ) greater than 0.2
8. **C ξ > 0.4** – The number of classes which have a class value (ξ) greater than 0.4

9. **C ξ > 0.5** – The number of classes which have a class value (ξ) greater than 0.5
10. **C ξ > 0.8** – The number of classes which have a class value (ξ) greater than 0.8
11. **Amalg. Classes** – The number of amalgamated classes (classes which combine more than one map unit)
12. **Max. Class Size** – The largest number of parent material classes in one unit. If this is greater than 1, and there are no amalgamated classes, this can arise due to the use of a simplified parent material classification where, from the onset, the class definitions are less specific than the fully detailed definition of soil parent material (Clayden and Hollis, 1984).
13. **% Unpredictable** – the percentage of the map which is unpredictable due to classes missing from the modelled map. (excluding urban areas, etc)
14. **Test** - the code by which this test is referred to in the text

Note: particularly successful or noteworthy tests may be circled in green, as shown above. For amalgamated tests, only the most successful amalgamation is reported.

5 DATA DICTIONARY METHODOLOGY

This chapter discusses methods of producing soil parent material maps from existing geological mapping. National and international parent material classifications are compared and used to create parent material maps from two sources of geological data: bedrock and surface geology. Initial analyses revealed consistent misclassification between classes, so two methods of classification simplification are investigated. The first simplifies the entire classification on the basis of lithological similarity. The second amalgamates commonly misclassified units. The results of this methodology are discussed and the methodology evaluated. Finally, recommendations for improvements to the initial map production are provided.

5.1 Introduction

The data dictionary methodology was developed to test the value of parent material maps, created by translation from geological maps. This type of translation from geology to parent material has been used elsewhere (Palmer et al., 2007; e-SOTER, 2008) and is perhaps the easiest way of generating a parent material map. This methodology tests the value of this approach.

One-to-one translational dictionaries between the geological classes and soil parent material classes were generated. The aim was to investigate methodologies to enable prediction of the soil parent material from geological data in regions without detailed soil mapping. Initial analyses showed that misclassification was widespread, and so two methods of classification simplification were designed and tested to overcome this problem.

5.1.1 Cartographic re-interpretation and translation

Traditional paper soil maps were occasionally reclassified at the time of publication in terms of land use capability and limitations (e.g. Hollis, 1978). Since the advent of geographic information systems (GIS), reclassification or interpretations of existing maps have become more common. Commonly, a simple translation table is used to convert the existing soil class to a class describing, for example, land trafficability, corrosivity or the vulnerability of a land to flooding. Recently a number of soil maps have been reclassified as parent material maps (RI USDA NRCS, 2009; USDA, 2002; BGR, 2004).

For detailed soil maps, such translations tend to be 1:1, where one soil class is attributed with one interpreted parent material class from a defined lookup table. For regional or national scale mapping, many parent materials may be included within a map unit. Because the linework for the interpreted parent material maps is derived from the original soil map, the act of reclassifying or interpreting the soil map can only remove linework, not add to it.

5.2 Parent material classifications

Parent material has been described using defined classifications and also using unconstrained descriptive text. Both national and international classifications of parent material exist. These different approaches are now discussed.

5.2.1 Descriptions of parent material (undefined classification)

A 1:5,000,000 scale map describing the parent material groups of Europe (BGR, 2004) has been created from an existing soil map. No distinct parent material classification was used, but rather, a description of the geological units and ages is provided for each soil unit. For example, the parent material description for one map unit is “Palaeozoic

sedimentary rocks, igneous and metamorphic rocks". The scale of this map gives rise to such broad descriptions. While some classes are more specific than others, most classes tend to include at least two quite different parent materials. The presence of such broad and undefined classes makes the map of limited use for soil or environmental modelling purposes, particularly if the link between parent material and specific soil series is important. Nevertheless, this is useful attribution for reference purposes, particularly where the identified parent material may be a superficial deposit which may have been omitted from existing geological mapping.

5.2.2 European Soil Bureau (ESB) classification

The European Soil Bureau (ESB) developed an explicit hierarchical and strongly lithological classification of soil parent material in the early 1990's (Lambert et al., 2003) for use at a nominal international scale of 1:1,000,000. This classification has been used in the SOTER (FAO, 1995) approach of characterising landscapes and soils. The e-SOTER project (e-SOTER, 2008) aims to test the application of these approaches in some study areas at a more detailed scale of 1:250,000. The parent material classification used in these projects (Table 12) is referred to as the ESB classification throughout this research.

There are 232 lithological classes within the ESB classification (Appendix 3). This classification allows a dominant and secondary parent material to be defined. Thus, in total there are 53,824 class combinations. This classification is hierarchical allowing classification to the level of available information. For example, in a case where little information exists, a parent material may be classed as a "consolidated clastic sedimentary rock (1000)", or where detailed information exists, a "calcareous sandstone (1211)" (Table 12).

As can be seen from Table 12, the ESB classification is strongly lithological, and, unlike the 1:5,000,000 map of Europe (BGR 2004), this classification does not provide

information on the age or stratigraphy of the units. Neither approach provides comment on the cohesion, consolidation or structure of the parent material.

Table 12 – Excerpt of the hierarchical ESB parent material classification (adapted from Lambert et al. (2003)).

Major Class level		Group level		Type level		Sub-type level	
1000	consolidated-clastic-sedimentary rocks	1100	psephite or rudite	1110	conglomerate	1111	pudding stone
				1120	breccia		
		1200	psammite or arenite	1210	sandstone	1211	calcareous sandstone
						1212	ferruginous sandstone
						1213	clayey sandstone
						1214	quartzitic sandstone orthoquartzite
						1215	micaceous sandstone
				1220	arkose		
				1230	graywacke	1231	feldspathic graywacke
		1300	pelite, lutite or argilite	1310	claystone / mudstone	1311	kaolinite
						1312	bentonite
				1320	siltstone		
		1400	facies bound rock	1410	flysch	1411	sandy flysch
						1412	clayey and silty flysch
						1413	conglomeratic flysch
				1420	molasse		

5.2.3 National Soil Resources Institute (NSRI) classification

While the geological age of the parent material has little effect on the resulting soil, the lithology and physical structure of the parent material exert strong controls on the soil and the properties of the near surface. The detailed National Soil Resources Institute (NSRI) soil parent material classification for England and Wales (Clayden & Hollis, 1984) addresses both these structural and lithological components of the parent material (Table 13 and Table 14).

Firstly, the presence or absence of certain structural features or diagnostic horizons within particular depths are used to classify the parent material into one of six broad parent material classes (PARENT), describing the broad physical nature of the substrate (Table 13). Once the soil parent material has been allocated to one of the PARENT classes, the lithological component (PM_LITH) of the classification is added to further define the parent material (Table 14).

Table 14 lists the PM_LITH classes which have been used in this research. A full description of this classification for all classes in England and Wales is provided in Appendix 2.

With the fully detailed NSRI classification, the parent material is described in terms of the physical nature of the substrate, and the lithology (or lithologies) of the parent material. The full, detailed PARENT + PM_LITH classification is given the name PARLITH. The PARLITH classes which occur in the study areas are listed in Table 15.

Table 13 – The broad PARENT component of the NSRI parent material classification

Parent Material	Descriptions
(A) Soils in peat	These are soils that meet both of the following criteria: (i) Either, more than 40 cm of organic material within the upper 80 cm of the profile, or more than 30 cm of organic material resting directly on bedrock or skeletal material. (ii) No superficial non-humose mineral horizons with a colour value of 4 or more that extend below 30 cm depth.
(B) Soils with a lithoskeletal substrate	These are mineral soils distinguished by the presence of a layer of angular material or coherent bedrock that is at least 15 cm thick and begins above and extends below 80 cm depth.
(C) Gravelly soils	These are mineral soils in which gravelly material extends from within 40 cm of the soil surface to at least 80 cm depth and which have no loamy or clayey surface layers more than 30 cm thick that contain less than 16 per cent stones by volume.
(D) Soils over gravel	Soils are described as being over gravel when they include both the following: (i) A gravelly layer more than 15 cm thick that starts above and extends below 80 cm depth. (ii) Either, at least 40 cm of superficial loamy or clayey material with less than 36 per cent stones by volume, or more than 30 cm of superficial loamy or clayey material with less than 16 per cent stones by volume.
(E) Soils in thick drift	These are mineral soils in Quaternary deposits at least 80 cm thick, that have no skeletal or textural contrasting gravelly layers extending below 80 cm depth. Soft pre-Quaternary material relatively uncontaminated by drift is absent from the upper 80 cm of the profile. Mineral soils in thin drift deposits which overlies organic layers that begin above and extend below 80 cm depth are also included in this parent material type.
(F) Soils in thin drift	These soils are distinguished by the presence within 80 cm depth, of either little altered soft pre-Quaternary material or a non-skeletal B horizon that passes conformably into pre-Quaternary material. Coherent bedrock may occur below 80 cm but is not present within this depth.

Table 14 – The PM_LITH component of the NSRI parent material classification

Note: this table describes the lithology and number of soil series which form over each parent material class. Only the classes which are found or predicted in the three study areas are listed. The full classification includes 96 classes. W: Worksop; N: Needwood Forest; Y: Yeovil; E&W: England and Wales.

PM_LITH	Parent Material Lithology	W	N	Y	E&W
Aa	sphagnum peat		1		2
Af	humified peat			1	7
Ba	acid crystalline rock				12
Bb	basic crystalline rock				18
Bc	ultrabasic crystalline rock				4
Bh	limestone	2		5	35
Bi	chalk			1	21
Bj	mudstone, shale or slate				2
Bk	siltstone, shale or slate				0
Bm	mudstone and sandstone or slate				20
Bn	siltstone and sandstone				3
Bo	sandstone	1	1		36
Bp	siltstone				4
Cg	sandstones, siltstones, mudstones or slate				2
Db	non-calcareous gravel		1		26
Ea	river alluvium	1	6	4	45
Ec	lake marl or tufa			1	4
Ee	non-calcareous colluvium			1	5
Ef	stoneless drift	2	2	1	57
Eg	chalky drift		1		40
Eh	drift with limestones			1	12
Ei	drift with siliceous stones	6	13	5	121
Fi	clay or soft mudstone	5	9	6	41
Fj	clay with interbedded limestone			1	4
Fm	loam (or soft sandstone, shale or siltstone)			2	9
Fq	sand or soft sandstone	1	1		24
Fu	loam or soft siltstone			2	3
Fy	soft shale or siltstone		1	3	8
	total	18	36	34	565

Table 15 - Parent material classes (PARLITH) which occur in the three study areas.

Parent Material Class
AaA3 - sphagnum peat (All other peat)
AfA3 - humified peat (All other peat)
BhB1 - limestone (Soils with lithoskeletal substrate)
BhB2 - limestone (Soils over lithoskeletal substrate)
BiB2 - chalk (Soils over lithoskeletal substrate)
BoB2 - sandstone (Soils over lithoskeletal substrate)
DbD1 - non-calcareous gravel (Soils over gravel)
EaE1 - river alluvium (Soils in thick drift)
EcE1 - lake marl or tufa (Soils in thick drift)
EeE1 - non-calcareous colluvium (Soils in thick drift)
EfE1 - stoneless drift (Soils in thick drift)
EgE1 - chalky drift (Soils in thick drift)
EhE1 - drift with limestones (Soils in thick drift)
EiE1 - drift with siliceous stones (Soils in thick drift)
FiF1 - clay or soft mudstone (Soils in thin drift passing to pre-Quaternary substrate)
FiF2 - clay or soft mudstone (Soils in soft pre-Quaternary material with no contrasting superficial drift)
FjF1 - clay with interbedded limestone (Soils in thin drift passing to pre-Quaternary substrate)
FmF1 - loam or soft sandstone, shale or siltstone (Soils in thin drift passing to pre-Quaternary substrate)
FqF1 - sand or soft sandstone (Soils in thin drift passing to pre-Quaternary substrate)
FuF1 - loam or soft siltstone (Soils in thin drift passing to pre-Quaternary substrate)
FyF1 - soft shale or siltstone (Soils in thin drift passing to pre-Quaternary substrate)

Note: For easy reference whilst reading, an identical table may be found on a fold out page in Appendix 2.

5.2.4 The use of parent material classifications by BGS

When the British Geological Survey (BGS) originally translated its 1:50,000 scale geological map to a parent material map according to the NSRI classification, some parent materials were found to have been attributed well, while there was great confusion with others (Palmer et al., 2007). It was concluded that a harmonisation between BGS and NSRI classifications would be beneficial.

However, the most recent version (v 4) of the BGS parent material map (Lawley, 2009) no longer employs the NSRI classification of parent material. Instead, this dataset has adopted a new suite of descriptive fields. It has also been attributed with a parent material according to the ESB classification. As the ESB classification places a strong emphasis on the lithology of the parent material, it is a simpler translation from geology than is the fully detailed NSRI classification. This may explain the adoption of the ESB classification. The simpler lithological component (PM_LITH) of the NSRI classification does not place such an explicit emphasis on the structure of the near surface, but some distinctions between drift, gravely and lithoskeletal soils are implicitly retained in this classification. As this new BGS dataset was released at the end of this research, there has not been time to examine it in any detail.

The main aim of this methodology was to investigate methods of translating traditional geological maps into soil parent material maps. In this chapter, both the NSRI and ESB classifications of parent material will be used to create parent material maps from existing geological data. The value of the resulting maps will be compared.

5.3 Assumptions

For this research, the following assumptions were made:

- That the 1:25,000 detailed soil maps accurately record the true distribution of soil type and soil parent material. Furthermore, that when complex units are described on these maps, the dominant soil type can be assumed to represent the whole unit.
- That the soil parent material is related to the mapped superficial and bedrock geology.

5.4 Methods

The datasets used in this method were:

- NSRI reference soil parent material maps (1:25,000)
- ESB reference soil parent material maps (1:25,000, derived from NSRI data)
- BGS bedrock and superficial geological maps (1:50,000) (GEOLOGY)

The NSRI and ESB parent material classifications were used to create parent material maps based on existing geological mapping. In the data dictionary methodology, only the lithological component of the NSRI classification (PM_LITH) was used. This was because geological maps do not tend to record the presence, within specific depths, of distinctive mineral substrates, organic matter or the physical nature of the top metre of regolith. Thus, it was not possible to accurately attribute the physical (PARENT) aspect of the parent material (Table 13).

The European Soil Bureau's (ESB) parent material classification (Lambert et al. 2002) was chosen as an alternative to the national NSRI classification for two main reasons. Firstly, the ESB classification is a European wide system. In the context of increasing pan-European initiatives and the harmonisation required for such projects, having a

classification which easily crosses borders is appealing. Secondly, this classification offers scope for dealing with units of mixed lithology through the use of dominant and secondary classes. This is important as many of the units on the geological maps are chronostratigraphic, rather than lithological. In practice, this means that it is possible to have multiple lithologies within a geological unit, which may be a closer representation of reality than a single lithological class for each geological class. Therefore, in the data dictionary methodology, the full ESB12 (dominant + secondary) classification has been used.

To allow comparison with later methodologies, which use the full PARLITH (PARENT+PM_LITH) classification, the map value metric (ψ_3) took into account the lower value of these broader ESB and PM_LITH classes by defining the number of classes (c) within each unit as the number of parent material classes of that unit, in the study area (see section 4.5.4). Thus, a comparison of class value between methodologies and classifications could be made.

The NSRI parent material was known from the 1:25,000 scale reference soil map in each study area. It was necessary to translate the NSRI classes to the ESB classification to create a reference ESB map. The geological units were translated to both the ESB and NSRI PM_LITH classes on the basis of their lithological descriptions. The modelled parent material maps were then tested, using metrics described in section 4.5, against reference parent material maps derived from detailed soil maps covering the study areas.

Three different approaches were investigated using detailed and simplified classifications of parent material. These were:

1. Detailed parent material classification (Approach 1)
2. Simplified parent material classification (Approach 2)
3. Guided amalgamation of parent material classes (Approach 3)

These approaches are now discussed in more detail.

5.4.1 Detailed parent material classifications – (Approach 1)

The geological map units were translated to soil parent material units using the ESB and NSRI PM_LITH classifications, which are compared in Table 16. On the basis of their lithological descriptions, each geological unit within the three study areas was classified to one parent material class in both the NSRI and ESB parent material classifications. The most lithologically similar class was chosen, and the relationships captured in translation tables. An extract from the geology to parent material translation table is provided in Table 17. This table shows the information from the BGS Lexicon which was used to allocate the units to the parent material codes to the identified NSRI and ESB classifications.

Table 16 – Approach 1 Parent Material Classifications

Classification	Members	Origin & Scale	Notes
PM_LITH	96	England & Wales 1:10,000 to 1:50,000	Lithological classification
ESB12	53,824 possible combinations of 232 parent material units	Europe 1:250,000 to 1:1,000,000	Hierarchical lithological classification with primary and secondary components

Parent material maps were created using these translation tables and the original geological datasets (Figure 20). Thus, the linework for these parent material maps was based entirely on the existing geological mapping. In some tests only the bedrock layer was used to create the parent material map (e.g. 4-6, see Table 19). This was to test whether or not the superficial layer added value to the parent material interpretation. In other tests, both bedrock and superficial layers were used, creating a surface geology layer.

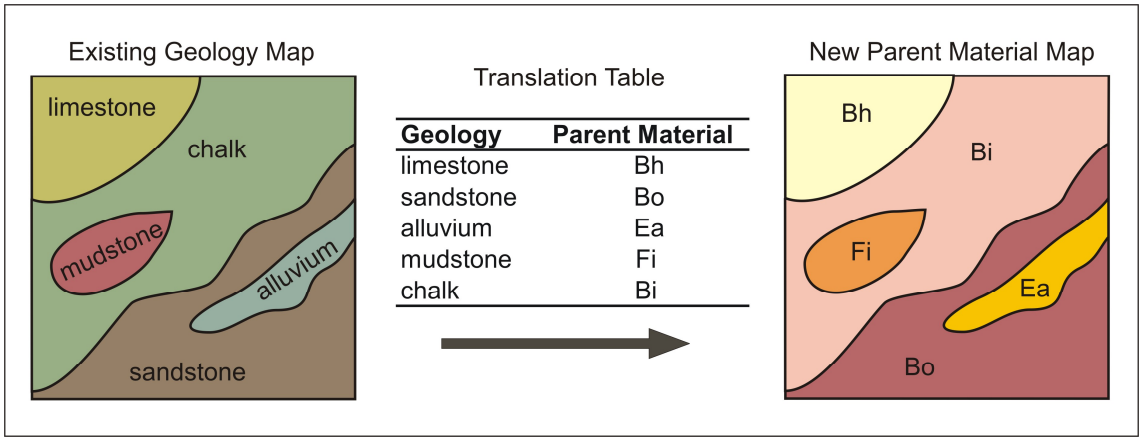


Figure 20 – Approach 1 –The translation from an existing geological map to a parent material map.
Note: the geology map was translated directly to the ESB classification as well as to the NSRI classification

5.4.2 Translational dictionaries

In summary, the following translational dictionaries were defined as part of this methodology:

- Superficial geology to NSRI PM_LITH parent material classification
- Bedrock geology to NSRI PM_LITH parent material classification
- Superficial geology to ESB parent material classification
- Bedrock geology to ESB parent material classification
- NSRI PM_LITH parent material classification to ESB parent material classification

Table 17 - Extract from the Approach 1 geology-to-parent material translation table

Lexicon code	Geological description from BGS Lexicon	NSRI code	ESB (dominant) code
CDF-CAMD	CADEBY FORMATION - CALCAREOUS MUDSTONE	Bj (mudstone, shale or slate)	1310 (claystone / mudstone)
EDT-CAMD	EDLINGTON FORMATION - CALCAREOUS MUDSTONE	Bj (mudstone, shale or slate)	1310 (claystone / mudstone)
UGS-CSDS	UPPER GREENSAND FORMATION - CALCAREOUS SANDSTONE	Bo (sandstone)	1211 (calcareous sandstone)
HCK-CHLK	HOLYWELL NODULAR CHALK FORMATION - CHALK	Bi (chalk)	2150 (chalk)
ZZCH-CHLK	ZIG ZAG CHALK FORMATION - CHALK	Bi (chalk)	2150 (chalk)
LECH-CHLK	LEWES NODULAR CHALK FORMATION - CHALK	Bi (chalk)	2150 (chalk)
NPCH-CHLK	NEW PIT CHALK FORMATION - CHALK	Bi (chalk)	2150 (chalk)
UGS-CHRT	UPPER GREENSAND FORMATION - CHERT	Cf (very hard siliceous stones)	2310 (chert, hornstone, flint)
CDF-DOLO	CADEBY FORMATION - DOLOMITE ROCK	Bh (limestone)	2120 (dolomite)
BLL-LMST	BEE LOW LIMESTONE FORMATION - LIMESTONE	Bh (limestone)	2110 (limestone)
BNLS-LMST	BEACON LIMESTONE FORMATION - LIMESTONE	Bh (limestone)	2110 (limestone)
CB-LMST	CORNBRASH FORMATION - LIMESTONE	Bh (limestone)	2110 (limestone)

5.4.2.1 Model analyses

Once the translational dictionaries were defined, the geological maps were reclassified according to the two soil parent material classifications. These were then compared with the two reference parent material maps. For each model, a range of summary statistics were calculated (see section 4.5), including the overall accuracy of the model (θ_1), an assessment of map value (ψ_3) and the value of each parent material class (ξ) identified in the study area.

While the results of this first approach will be discussed in more detail later, it is helpful at this stage to note the poor performance of this first approach. Extensive map disagreement was found using both parent material classifications, in all study areas. This lack of success led to the development of methods of classification simplification which are now described.

5.4.3 Simplified parent material classifications (Approach 2)

Because of the large number of potential parent material classes in both the PM_LITH and ESB12 classifications (96 and 53,824, respectively), lithological similarity between the parent material classes was identified as an issue contributing to consistent misclassification in Approach 1. By combining parent material units with lithological similarity, the NSRI classification was reduced from 96 classes to 27 (Appendix 2). Likewise, the four levels of increasing lithological simplification were used from the hierarchical ESB classification resulting in 232 subtypes, 152 types, 50 groups and 9 major classes, respectively (Table 12). These simplified classifications are compared in Table 18 and this approach described in Figure 21.

Table 18 – Approach 2 Simplified Parent Material Classifications
Note: * indicates an Approach 1 (detailed) classification, presented for comparison

Classification	Members	Area of Origin	Notes
PM_LITH *	96	England and Wales	Only using lithological component
NSRI simplified	27	England and Wales	Aggregates similar lithologies
ESB12 *	53,824 possible combinations of 232 parent material units	Europe	Hierarchical lithological classification with primary and secondary components
ESB subtype	232	Europe	Derived from ESB12 (primary component only)
ESB type	152		
ESB group	50		
ESB major group	9		

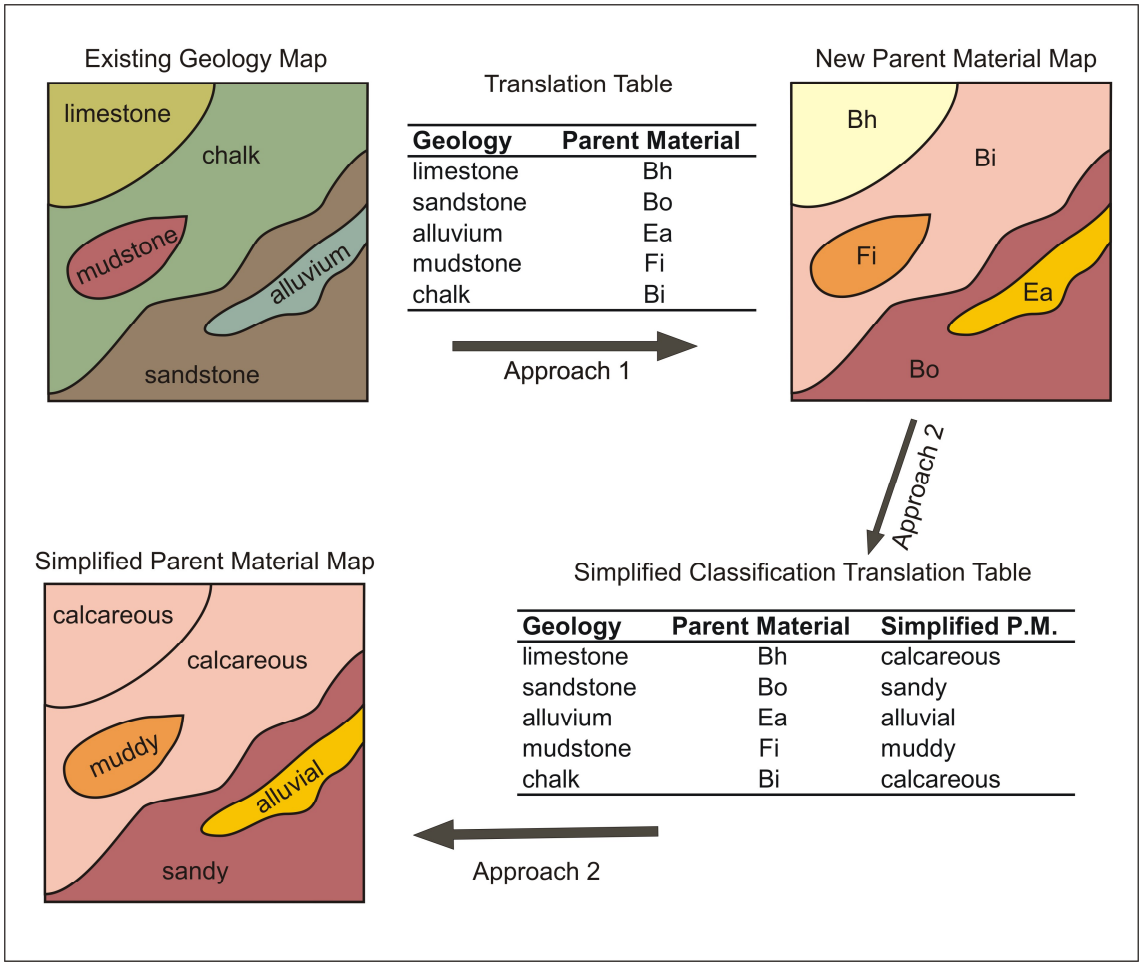


Figure 21 – Approach 2 –The translation from an existing geological map to a parent material map, and then to a parent material map, simplified on the basis of lithology.
Note: the geology map was translated directly to the ESB classification as well as to the NSRI classification. Simplification was implicit in the hierarchy of the ESB classification (Table 18).

As with Approach 1, the geological and soil maps were translated to the simplified parent material classifications, compared, and metrics generated. The simplified classification (Approach 2) produced some maps of higher value than Approach 1, however, significant misclassification remained between parent material units and some unnecessary reduction in class detail. Additionally, there remained extensive map disagreement between the modelled and reference maps. Therefore an alternative method of classification simplification was tested, that of selective, or guided amalgamation of commonly misclassified parent material classes (Approach 3).

5.4.4 Guided amalgamation of parent material units (Approach 3)

For the guided amalgamation approach, the 96-member NSRI PM_LITH classification, and the 232-class ESB subtype classification were used. The full ESB12 classification with 53,824 members was not used as the level of information gained from the secondary classes did not provide sufficient improvements to warrant the complexity of the full classification. This was because, in most cases, the dominant and secondary parent materials were the same. Additionally, with class amalgamation, the ability to describe units of mixed lithology is gained, but in a more flexible manner.

Following the creation of these initial parent material maps, spatial analysis was used to guide the amalgamation of frequently misclassified, yet broadly similar, units together, forming wider parent material classes with higher levels of spatial agreement. This process is described in Figure 22.

Classes could be amalgamated on the basis of:

- Lithological similarity
- Structural / physical similarity
- Extensive misclassification which may represent a consistent difference in mapping approaches between geologists and soil scientists.

Approaches similar to class amalgamation have been employed before, for example, as a cartographic tool for representing areas where there is complex heterogeneity within a mapping unit. The mapping units on the National Soil Map include multiple soil series, yet these are grouped together as the linework needed to delineate each soil series would be too complex to display on a 1:250,000 scale map. Indeed, there are even mixed soil units on the 1:25,000 scale maps which have been used to create the reference parent material maps for this research. Because the 1:50,000 scale geological maps are acting as the predictor of parent material, it is probable that the broader classes will encompass multiple soil parent materials. This justifies the use of mixed or amalgamated classes.

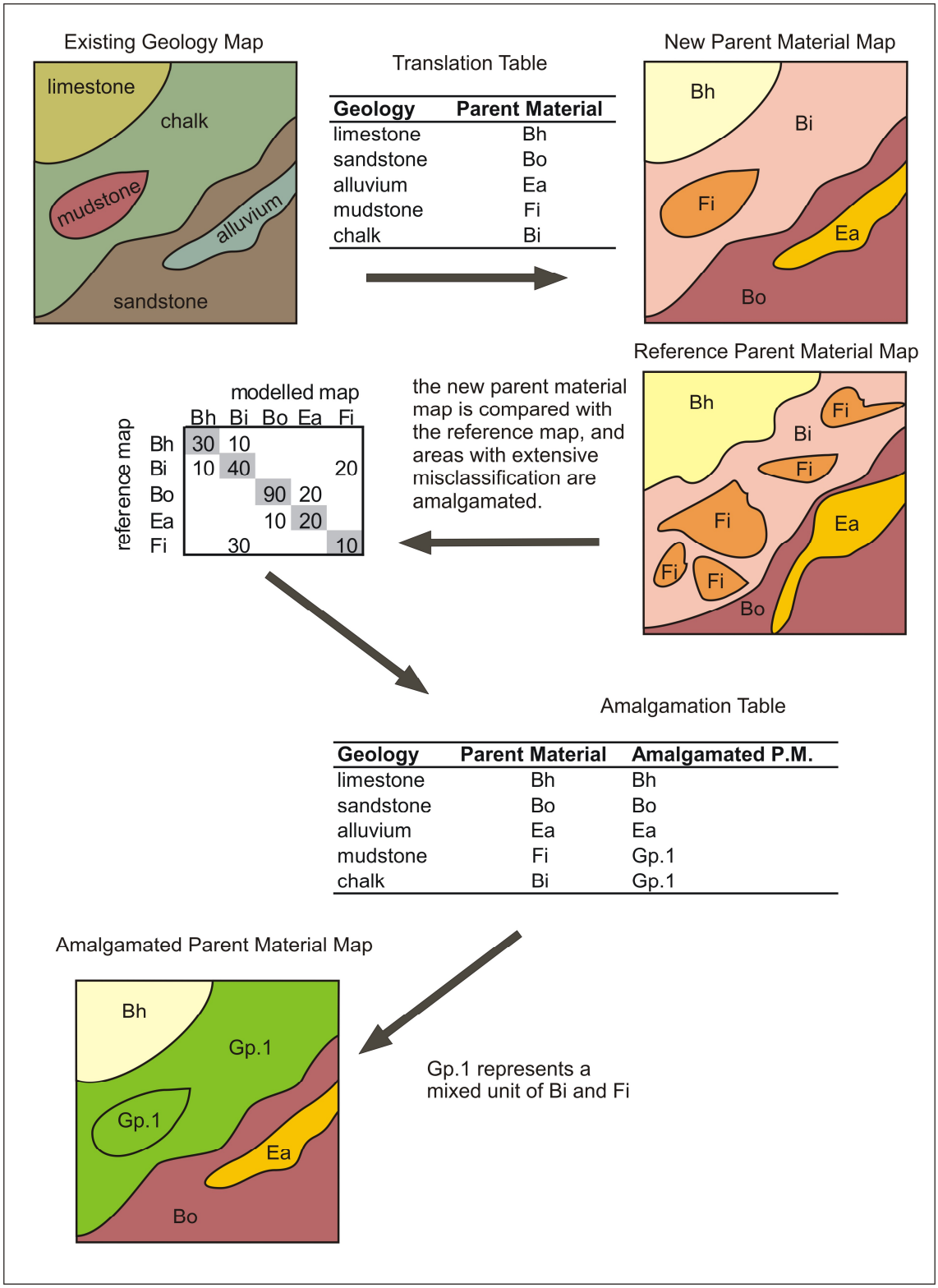


Figure 22 – Approach 3 -The process of guided amalgamation
Note: the geology map was translated directly to the ESB classification as well as to the NSRI classification. Class amalgamation was carried out in an identical method for each classification.

Numerous class amalgamation combinations were undertaken for each test. The aim was to maximise the map value (ψ_3), while respecting the amalgamation guidelines. Only the tests resulting in the most valuable maps have been reported for each classification.

A possible alternative approach to correct misclassifications would be to move the incorrectly classified parent material units to the class which is in agreement with the reference parent material map. A similar approach to this will be investigated later in the research, but at this juncture, the two units have instead been amalgamated.

5.5 Data dictionary methodology results

The results from the data dictionary methodology are summarised in Table 19 to Table 21. Particularly successful tests are circled for each parent material classification.

Table 19 - Data dictionary results for Worksop
Note: Presenting results for parent material maps using the PM_LITH and ESB classifications. A1: full detailed classification; A2: simplified classification; A3: amalgamated classes; (surf): surface geology input; (bed): bedrock geology input. For explanations of the headings, see Table 9, page 70.

Worksop Method		κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
PM_LITH (surf.)	(A1)	0.27	**0.40**	**0.28**	10	3	2	1	1	**1**	-	**2**	57%	W1
PM_LITH (surf.)	(A2)	0.27	**0.42**	**0.22**	9	2	2	1	1	**1**	-	**3**	50%	W2
PM_LITH (surf.)	(A3)	0.74	**0.83**	**1.18**	7	4	3	3	3	**2**	2	**4**	2%	W3
PM_LITH (bed.)	(A1)	0.28	**0.42**	**0.29**	9	2	2	1	1	**1**	-	**2**	57%	W4
PM_LITH (bed.)	(A2)	0.27	**0.43**	**0.22**	8	2	2	1	1	**1**	-	**3**	50%	W5
PM_LITH (bed.)	(A3)	0.78	**0.86**	**1.25**	5	4	3	3	3	**2**	3	**4**	0%	W6
ESB12 (surf.)	(A1)	0.00	**0.03**	**0.00**	11	2	-	-	-	**-**	-	**2**	50%	W7
ESB subtype (surf.)	(A2)	0.11	**0.17**	**0.05**	11	4	2	1	1	**-**	-	**4**	31%	W8
ESB type (surf.)	(A2)	0.11	**0.17**	**0.05**	11	4	2	1	1	**-**	-	**4**	31%	W9
ESB group (surf.)	(A2)	0.39	**0.53**	**0.39**	9	5	3	2	2	**1**	2	**4**	31%	W10
ESB major group (surf.)	(A2)	0.40	**0.59**	**0.35**	4	3	2	2	2	**1**	4	**8**	0%	W11
ESB subtype (surf.)	(A3)	0.70	**0.80**	**0.78**	6	4	3	3	3	**2**	3	**7**	0%	W12
ESB12 (bed.)	(A1)	0.00	**0.02**	**0.00**	11	1	-	-	-	**-**	-	**2**	88%	W13
ESB subtype (bed.)	(A2)	0.12	**0.18**	**0.06**	7	3	2	1	1	**-**	-	**4**	33%	W14
ESB type (bed.)	(A2)	0.12	**0.18**	**0.06**	7	3	2	1	1	**-**	-	**4**	33%	W15
ESB group (bed.)	(A2)	0.41	**0.56**	**0.43**	6	3	3	2	2	**1**	1	**4**	33%	W16
ESB major group (bed.)	(A2)	0.42	**0.61**	**0.37**	3	2	2	2	2	**1**	3	**7**	33%	W17
ESB subtype (bed.)	(A3)	0.74	**0.83**	**0.89**	5	3	3	3	3	**2**	2	**5**	2%	W18

Table 20 - Data dictionary results for Needwood Forest

Note: Presenting results for parent material maps using the PM_LITH and ESB classifications. A1: full detailed classification; A2: simplified classification; A3: amalgamated classes; (surf): surface geology input; (bed): bedrock geology input. For explanations of the headings, see Table 9, page 70.

Needwood Forest														
Method		κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
PM_LITH (surf.)	(A1)	0.04	**0.04**	**0.04**	12	**3**	3	2	2	-	-	**2**	95%	N1
PM_LITH (surf.)	(A2)	0.14	**0.26**	**0.23**	9	**4**	3	3	3	-	-	**2**	69%	N2
PM_LITH (surf.)	(A3)	0.62	**0.95**	**1.43**	6	**4**	4	4	4	**1**	2	**7**	2%	N3
PM_LITH (bed.)	(A1)	0.00	**0.00**	**0.00**	11	**1**	-	-	-	-	-	**2**	100%	N4
PM_LITH (bed.)	(A2)	0.01	**0.29**	**0.09**	9	**2**	2	1	1	-	-	**2**	71%	N5
PM_LITH (bed.)	(A3)	0.16	**0.96**	**0.45**	5	**2**	2	1	1	**1**	2	**7**	2%	N6
ESB12 (surf.)	(A1)	0.13	**0.24**	**0.08**	15	**1**	1	1	1	-	-	**2**	71%	N7
ESB subtype (surf.)	(A2)	0.13	**0.24**	**0.10**	11	**3**	2	2	2	-	-	**4**	64%	N8
ESB type (surf.)	(A2)	0.13	**0.24**	**0.10**	11	**3**	2	2	2	-	-	**4**	64%	N9
ESB group (surf.)	(A2)	0.13	**0.25**	**0.18**	8	**5**	3	3	3	-	3	**4**	57%	N10
ESB major group (surf.)	(A2)	0.17	**0.33**	**0.18**	4	**3**	3	2	2	-	4	**8**	0%	N11
ESB subtype (surf.)	(A3)	0.11	**0.90**	**0.42**	7	**2**	2	2	2	**1**	1	**9**	7%	N12
ESB12 (bed.)	(A1)	0.01	**0.29**	**0.06**	10	**1**	1	1	1	-	-	**2**	71%	N13
ESB subtype (bed.)	(A2)	0.01	**0.29**	**0.07**	7	**2**	2	1	1	-	-	**4**	71%	N14
ESB type (bed.)	(A2)	0.01	**0.29**	**0.07**	7	**2**	2	1	1	-	-	**4**	71%	N15
ESB group (bed.)	(A2)	0.01	**0.29**	**0.07**	6	**2**	2	1	1	-	1	**4**	70%	N16
ESB major group (bed.)	(A2)	0.00	**0.30**	**0.03**	3	**1**	1	1	1	-	2	**8**	70%	N17
ESB subtype (bed.)	(A3)	0.26	**0.98**	**0.39**	4	**2**	2	1	1	**1**	1	**9**	0%	N18

Table 21 - Data dictionary results for Yeovil

Note: Presenting results for parent material maps using the PM_LITH and ESB classifications. A1: full detailed classification; A2: simplified classification; A3: amalgamated classes; (surf): surface geology input; (bed): bedrock geology input. For explanations of the headings, see Table 9, page 70.

Yeovil Method		κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
PM_LITH (surf.)	(A1)	0.11	**0.13**	**0.15**	19	3	3	3	3	-	-	2	83%	Y1
PM_LITH (surf.)	(A2)	0.11	**0.14**	**0.23**	17	5	4	4	4	1	-	3	72%	Y2
PM_LITH (surf.)	(A3)	0.62	**0.76**	**1.39**	11	6	6	6	5	1	4	6	1%	Y3
PM_LITH (bed.)	(A1)	0.06	**0.07**	**0.04**	18	2	2	2	2	-	-	2	92%	Y4
PM_LITH (bed.)	(A2)	0.11	**0.14**	**0.09**	17	4	3	3	3	-	-	3	73%	Y5
PM_LITH (bed.)	(A3)	0.66	**0.83**	**0.97**	8	4	4	4	4	1	3	7	1%	Y6
ESB12 (surf.)	(A1)	0.11	**0.13**	**0.13**	24	4	3	3	3	-	-	2	77%	Y7
ESB subtype (surf.)	(A2)	0.37	**0.48**	**0.70**	15	7	5	5	5	-	-	4	13%	Y8
ESB type (surf.)	(A2)	0.37	**0.48**	**0.67**	13	7	5	5	5	-	2	5	13%	Y9
ESB group (surf.)	(A2)	0.32	**0.49**	**0.26**	8	7	3	3	3	-	5	5	0%	Y10
ESB major group (surf.)	(A2)	0.47	**0.75**	**0.38**	4	3	3	3	3	1	3	9	0%	Y11
ESB subtype (surf.)	(A3)	0.54	**0.69**	**1.03**	13	6	5	5	5	1	2	5	4%	Y12
ESB12 (bed.)	(A1)	0.11	**0.14**	**0.12**	19	4	3	3	3	-	-	2	77%	Y13
ESB subtype (bed.)	(A2)	0.38	**0.50**	**0.67**	12	5	5	5	5	-	-	4	0%	Y14
ESB type (bed.)	(A2)	0.38	**0.50**	**0.66**	11	5	5	5	5	-	1	5	0%	Y15
ESB group (bed.)	(A2)	0.31	**0.51**	**0.26**	8	3	3	3	3	-	3	5	0%	Y16
ESB major group (bed.)	(A2)	0.20	**0.70**	**0.18**	3	2	2	1	1	1	2	11	0%	Y17
ESB subtype (bed.)	(A3)	0.65	**0.82**	**0.91**	7	4	4	4	4	1	2	9	0%	Y18

5.6 Discussion of the data dictionary methodology

The following discussions examine and compare the results of the different approaches and classifications used in the data dictionary methodology. Comparisons are drawn between the fully detailed ESB and NSRI classifications, finding that both produce poor results with extensive misclassification (Approach 1). The two classification simplification approaches employed to correct this misclassification are evaluated. The simplified classifications (Approach 2) achieve only marginal improvement in the map value (ψ_3), but the use of guided amalgamation (Approach 3) is found to greatly improve map value. The successes of the translations based on the two geological inputs (bedrock & surface geology) are compared. Parent materials classification success is discussed. Finally, the overall success of the methodology is reviewed.

5.6.1 Fully detailed parent material classifications (Approach 1)

For the analysis of Approach 1 (identified with A1 in the method column in Table 19 to Table 21), please refer to Tests 1, 4, 7 and 13 for each study area (identified by the prefix W, N and Y for Worksop, Needwood Forest and Yeovil, respectively).

5.6.1.1 Comparing the fully detailed European and national parent material classifications

The European ESB classification offers a wide range of discrete, lithological units in a hierarchical structure, which allows simple linking to lithological or chronostratigraphic geological maps with minor lithological variation within the map units. The full ESB12 classification allows for a dominant and secondary class to be specified, which enables limited mixed lithology units to be characterised.

The NSRI PM_LITH classification has fewer classes of lithology than the ESB classification, and is not hierarchical, but offers both single (e.g. Bh – Limestone) or

previously defined complex classes (e.g. Fr – Sand with interbedded limestone). However there is significantly less flexibility with these pre-defined complex classes than with the ESB12 approach. Nevertheless, the ESB principal of dominant and secondary classes might be usefully incorporated into parent material maps using the NSRI classification, should there be the need to describe a range of parent material in one mapping unit. Indeed this principal offers scope for describing more than two classes and should be investigated further.

Using the fully detailed parent material classifications, neither the European (ESB12, Tests 7 and 13), nor the national (PM_LITH, Tests 1 and 4) classification produced parent material maps with high map values (ψ_3) in any of the study areas. There were few effective classes (C_e) in the translations. This is particularly evident in the complex ESB12 classification where as few as 1 in 15 effective classes were found in both the reference and predicted maps (Test N7).

While broadly similar spatial patterns can be seen between the reference and predicted maps, there is limited actual agreement, as shown by the extent of the red 'Incorrect' class on Figure 23 (c), for the Yeovil area (Test Y1). This is due both to misclassification of geological units and to the difference in scale and detail, between the 1:25,000 reference map (shown in Figure 23 (a)), and the 1:50,000 geological map used to predict the parent material.

In this methodology, a large number of units are typically predicted which do not actually occur on the reference soil maps (e.g. Bj and Cg in Test N1, see Figure 24). This misclassification occurs because each geological unit was classified to a parent material unit with no knowledge of the actual units on the reference soil parent material map. In this case, there is significant misclassification between the predicted dark red Cg (sandstones, siltstones, mudstones of slate – gravely soils; Figure 24b), and actual light blue Ei (drift with siliceous stones; Figure 24a). These two classes are not identical, but are similar.

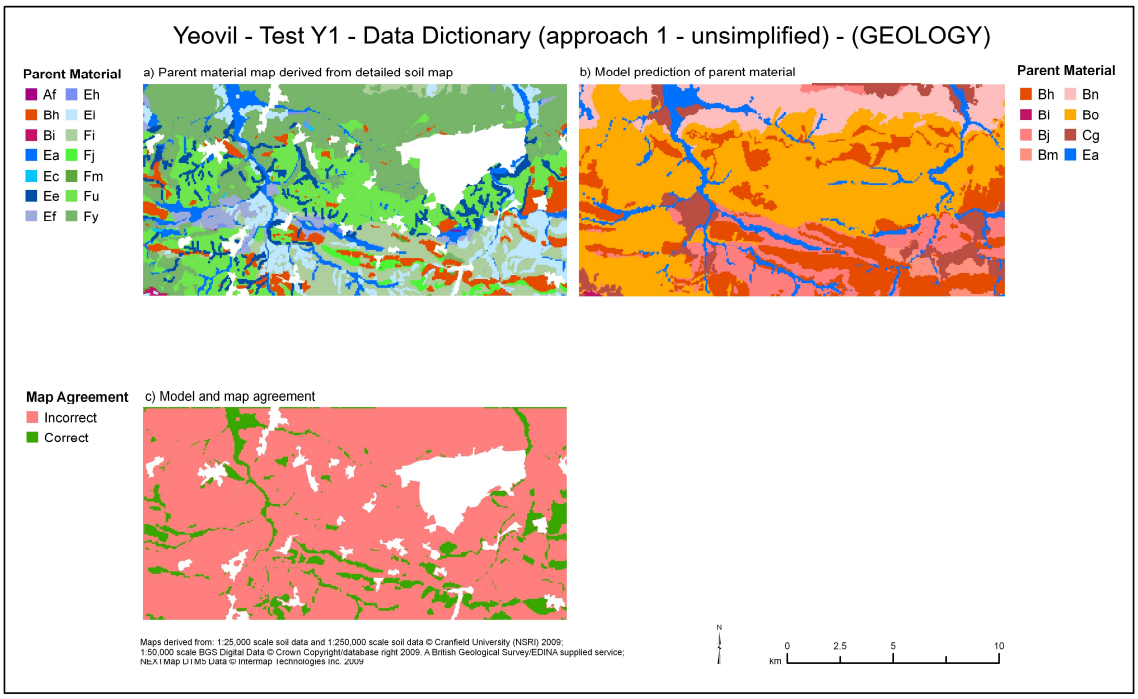


Figure 23 - Test Y1 maps (Approach 1)
Input: GEOLOGY (surface); Classification: NSRI PM_LITH; $\Psi_3 = 0.15$; $\theta_1 = 0.13$ $C_e = 3$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

The NSRI PM_LITH classification clearly outperformed the ESB12 classification in Worksop (NSRI: W1; $\Psi_3 = 0.28$ versus ESB: W7; $\Psi_3 = 0.00$), but in Yeovil and Needwood Forest the differences were negligible as both classifications produced very poor results.

5.6.1.2 Issues with the complexity of the classifications

The fully detailed, two-class ESB12 classification (Test Y7; $\Psi_3 = 0.13$) performed poorly, as the class complexity resulted in very little agreement between the model and the reference parent material map. Indeed, agreement only occurred where the dominant and secondary classes were the same.

The low number of effective classes (4 out of 24 classes, Test Y7 and 1 out of 15 classes, Test N7), indicate that the ESB12 classification is overly complex for this purpose and produces too many classes, and too many possible class combinations for

the level of detail and information held on the geological map. Similar issues exist for the PM_LITH classification, where, for example in Test Y1, there are only 3 effective classes out of 19 in total. This produces extensive incorrect prediction of parent material (e.g. see all the red on Figure 23c). The widespread misclassification of the geological map to parent material units not found on the reference maps also increases the total number of classes.

Approach 2 attempted to address these issues of poor classification by simplifying the parent material classification. Lithologically similar parent materials were combined into broader classes in an attempt to increase classification success.

5.6.2 Simplified parent material classifications (Approach 2)

For the analysis of Approach 2 (identified with A2 in the method column in Table 19 to Table 21), please refer to Tests 2, 5, 8-11 and 14-17 for each study area (identified by the prefix W, N and Y for Worksop, Needwood Forest and Yeovil, respectively). These are the simplified classifications.

The full parent material classifications were simplified on the basis of lithological criteria, with no reference to the spatial agreement from the initial tests in Approach 1. The ESB classification was simplified at four levels using the hierarchical structure for only the dominant class (Table 12). The NSRI PM_LITH classification is not hierarchical, so only one level of simplification on the basis of lithology was applied (Appendix 2).

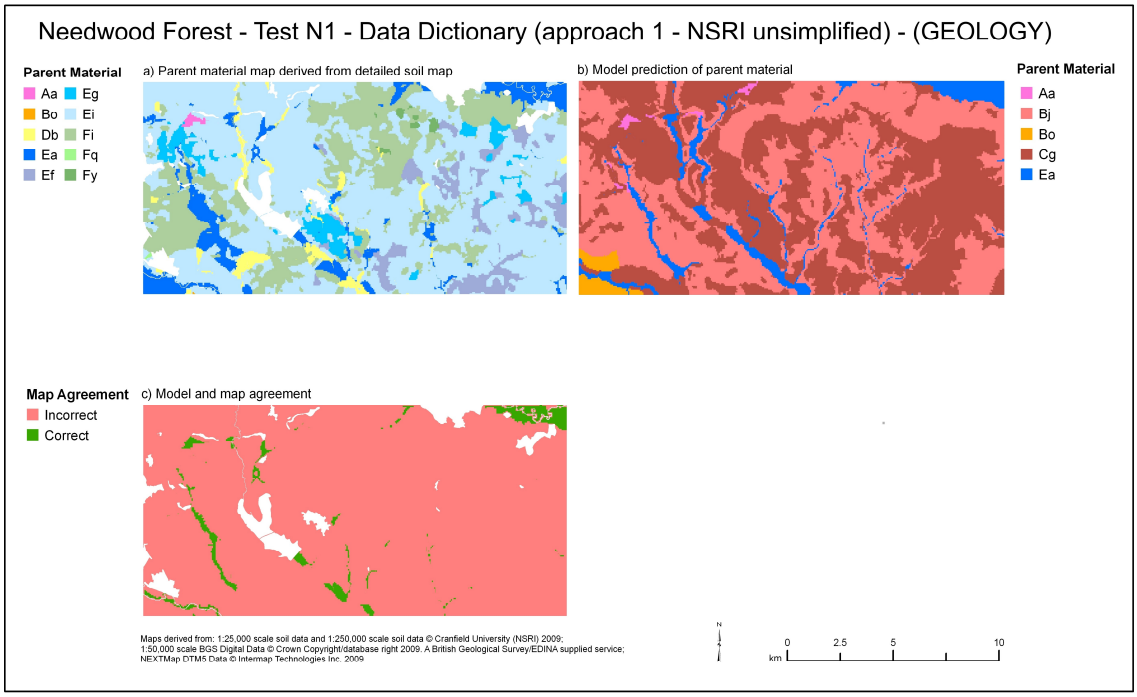


Figure 24 - Test N1 maps (Approach 1)
Input: GEOLOGY (surface); Classification: NSRI PM_LITH; $\Psi_3 = 0.04$; $\theta_1 = 0.04$ $C_e = 3$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

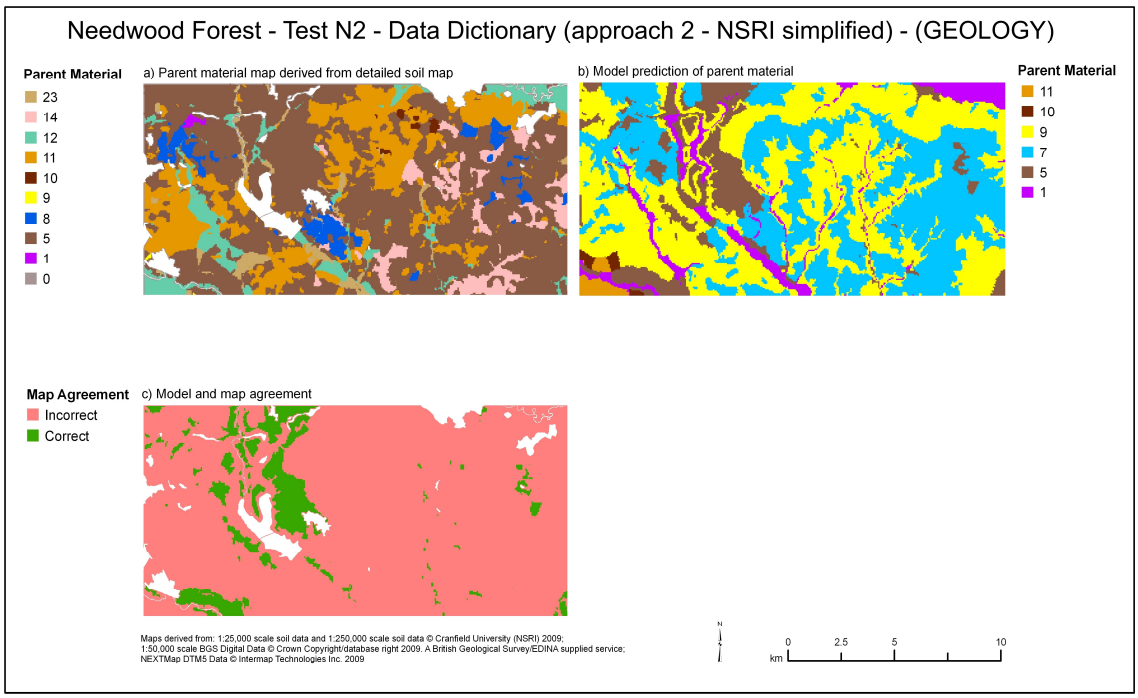


Figure 25 – Test N2 maps (Approach 2)
Input: GEOLOGY (surface); Classification: NSRI simplified; $\Psi_3 = 0.23$; $\theta_1 = 0.26$ $C_e = 4$
A larger version is available in Appendix 4. NSRI simplified codes are described in Appendix 2.

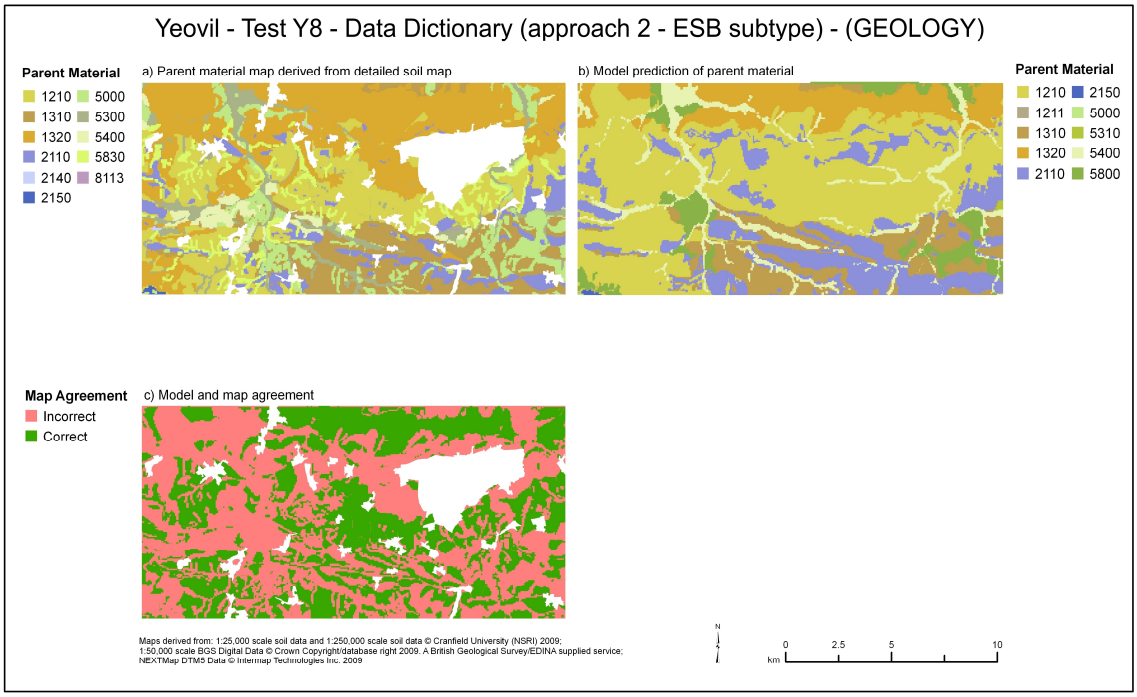


Figure 26 - Test Y8 maps (Approach 2 – ESB subtype)
Input: GEOLOGY (surface); Classification: ESB subtype; $\Psi_3 = 0.70$; $\theta_1 = 0.48$ $C_e = 7$
A larger version is available in Appendix 4. ESB codes are described in Appendix 3.

The value (Ψ_3) of the resulting parent material map almost always increased with the simplification of the classifications over the initial tests. Using the NSRI classification, the improvement in Needwood Forest can be seen by the increase in the extent of green (agreement) between Maps N1 (Figure 24c) and N2 (Figure 25c). Simply removing the secondary class from the ESB12 classification always improved map value, and occasionally there was a dramatic improvement, for example, ESB12: Y7 $\Psi_3 = 0.13$ versus ESB (subtype): Y8 $\Psi_3 = 0.70$ (Figure 26). The improvement in agreement can be seen clearly when ESB subtype (W8, Figure 27) is simplified to group level (W10, Figure 28). W10 achieved the most effective classes yet extensive misclassifications still remain between unconsolidated deposits (5000) and sandstone (1210).

The two exceptions where classification simplification did not bring about improvement were Tests W2 and N17. In W2, some classes were unnecessarily simplified, leading to a lower map value. In N17, where the ESB classification is simplified to the very general major group level, there also is an unnecessary loss in classification detail. The less drastic ESB simplifications (subtype to group level (Tests N14 to N16)) all brought

about a very small improvement over the ESB12 test (N13) by increasing the number of effective classes through classification simplification.

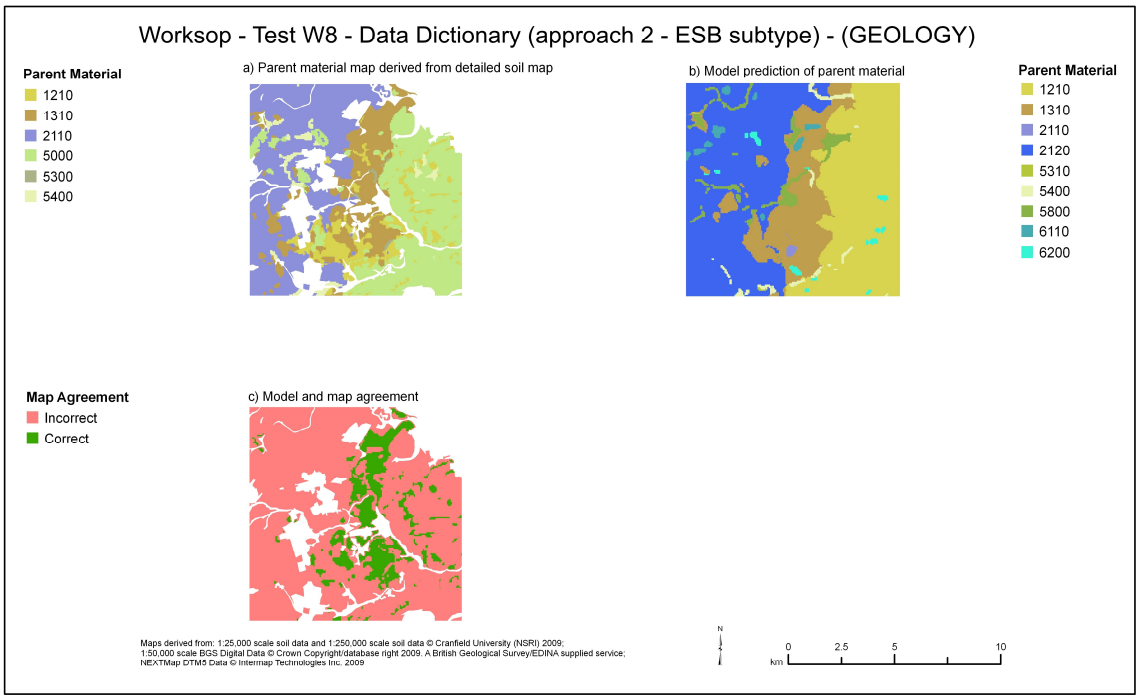


Figure 27 – Test W8 maps (Approach 2 – ESB subtype)
Input: GEOLOGY (surface); Classification: ESB subtype; $\Psi_3 = 0.05$; $\theta_1 = 0.17$ $C_e = 4$
A larger version is available in Appendix 4. ESB codes are described in Appendix 2.

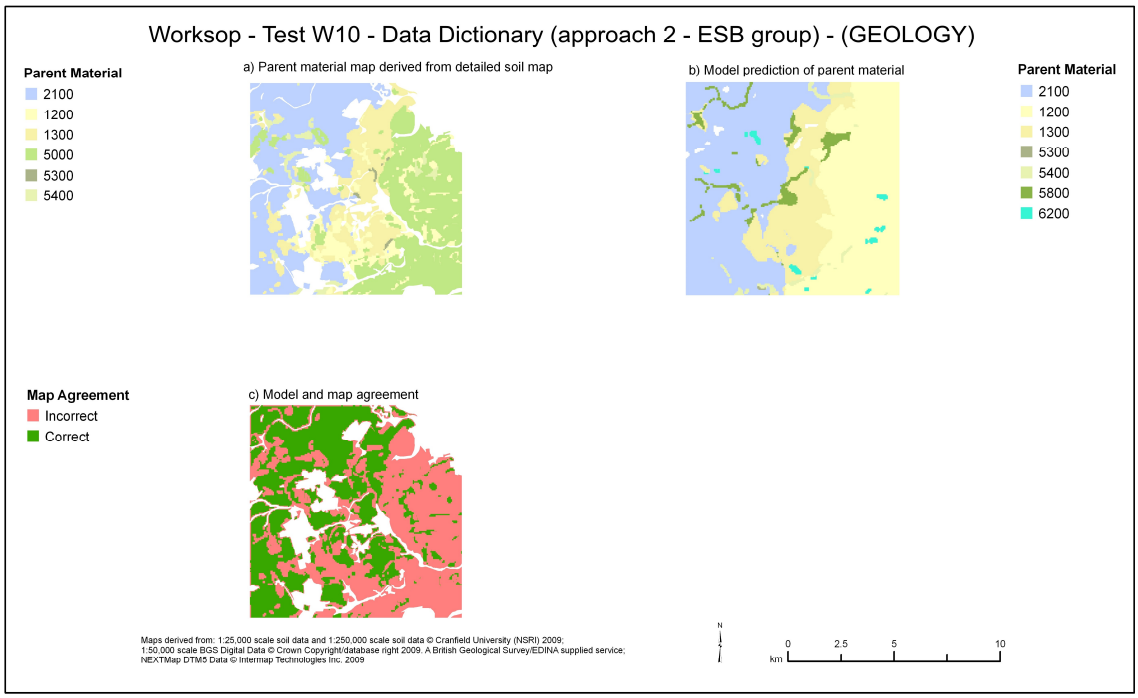


Figure 28 – Test W10 maps (Approach 2 – ESB group)
Input: GEOLOGY (surface); Classification: ESB group; $\Psi_3 = 0.39$; $\theta_1 = 0.53$ $C_e = 5$
A larger version is available in Appendix 4. ESB codes are described in Appendix 2.

Nevertheless, while there were some improvements, simplification of the whole classification did not produce maps which showed considerable improvements over Approach 1. Because of generic simplification, many classes were unnecessarily combined, lowering the weighted class values (ω) more than would have been implemented given a case by case approach to class amalgamation. Therefore, Approach 3 was undertaken to selectively amalgamate classes, in order to keep as much class detail as possible, only amalgamating classes when necessary. It was hypothesised that this third approach would achieve more valuable maps.

5.6.3 Guided amalgamation of parent material units (Approach 3)

For the analysis of Approach 3 (identified with A3 in the method column in Table 19 to Table 21), please refer to Tests 3, 6, 8, 12 and 18 for each study area (identified by the prefix W, N and Y for Worksop, Needwood Forest and Yeovil, respectively). These are the amalgamated classifications.

The guided amalgamation approach always outperformed the lithological grouping used in Approach 2 (compare the extent of correct prediction between Figure 29 and Figure 30). This success was predominantly due to the inherent flexibility in this classification simplification approach. Using the guided amalgamation approach, classes which were performing well (achieving high class values (ξ)) could be left with full class detail, while those classes which were performing poorly could be grouped to maximise correct prediction of a diluted parent material class. These broader, amalgamated classes tend to be lithologically similar, although occasionally, parent material classes have been grouped with reference to their origin, for example, drift deposits of differing lithologies may be amalgamated. Such amalgamations were performed as the geological mapping did not differentiate between lithological differences in the drift mapping.

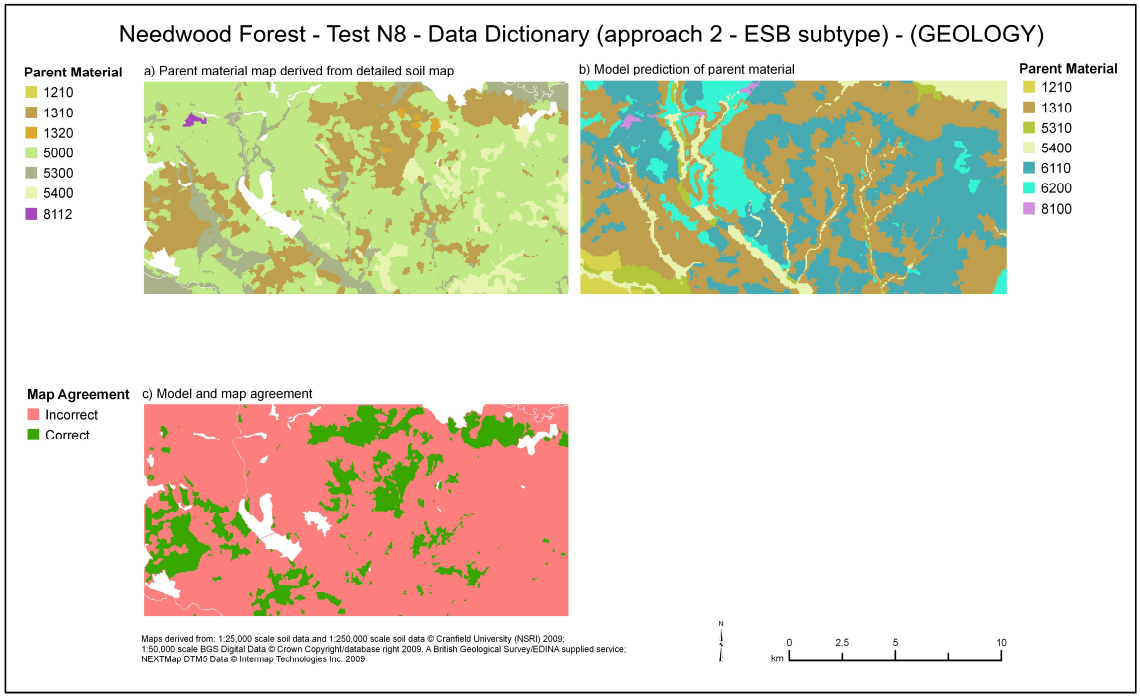


Figure 29 – Test N8 maps (Approach 2 – ESB subtype)
Input: GEOLOGY (surface); Classification: ESB subtype; $\Psi_3 = 0.10$; $\theta_1 = 0.24$ $C_e = 3$
A larger version is available in Appendix 4. ESB codes are described in Appendix 2.

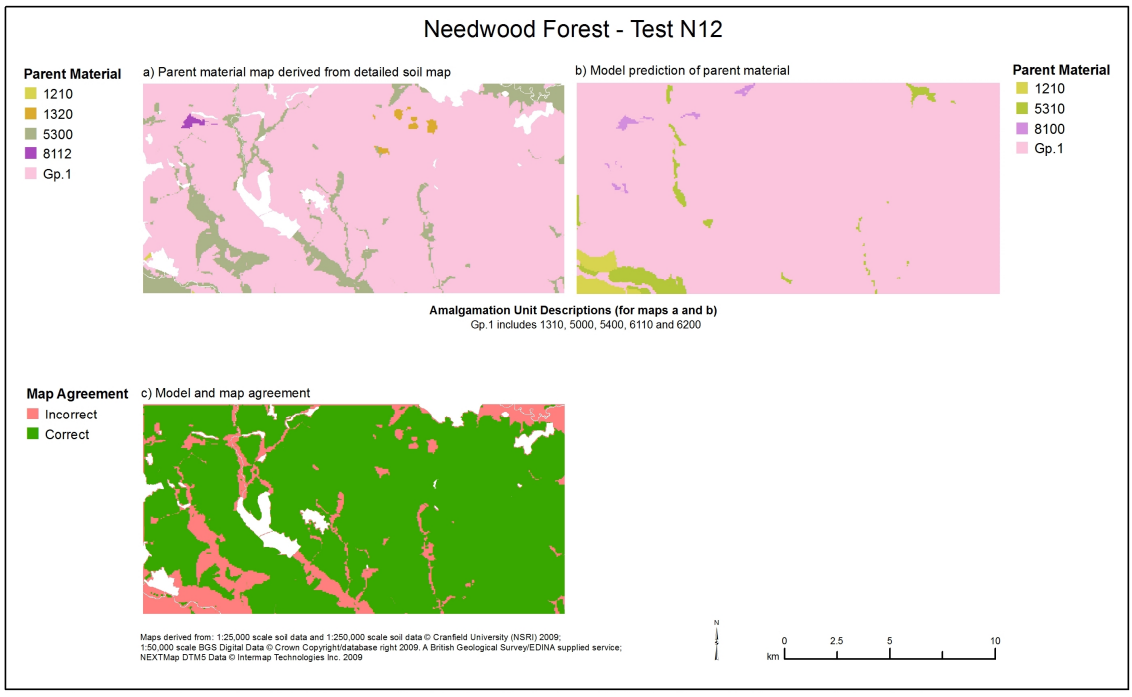


Figure 30 – Test N12 maps (Approach 3 – ESB amalgamated)
Input: GEOLOGY (surface); Classification: amalgamated ESB subtype ; $\Psi_3 = 0.42$; $\theta_1 = 0.90$ $C_e = 2$
A larger version is available in Appendix 4. ESB codes are described in Appendix 2.

A comparison of model results derived from the straight 1:1 translation from Approach 1 (W1: $\Psi_3 = 0.28$, Figure 31) and Approach 3's guided amalgamation (W3: $\Psi_3 = 1.18$, Figure 32) clearly demonstrate the advantage of amalgamating parent material classes to achieve higher levels of agreement (green). In this case, two amalgamated classes were created, one with two members, one with three (Figure 32). These amalgamations dealt with the misclassification of similar units and led to a much more valuable map.

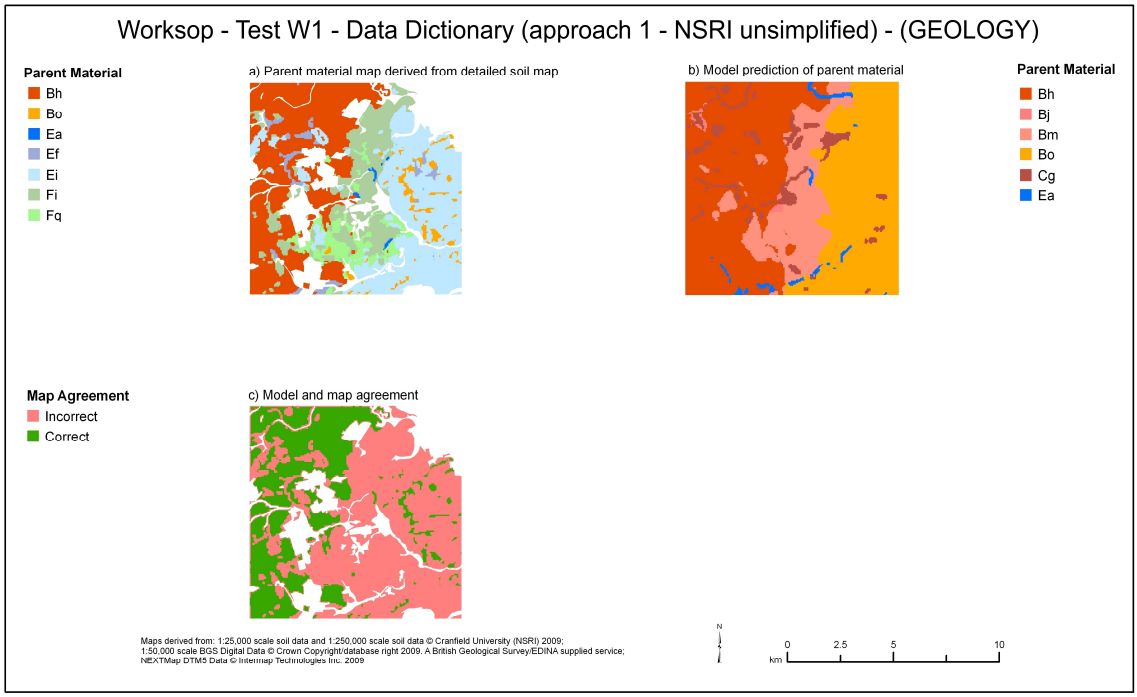


Figure 31 - Test W1 maps (Approach 1)
Input: GEOLOGY (surface); Classification: NSRI PM_LITH; $\Psi_3 = 0.28$; $\theta_1 = 0.40$ $C_e = 3$
A larger version is available in Appendix 4. NSRI PM_LITH codes are described in Appendix 2.

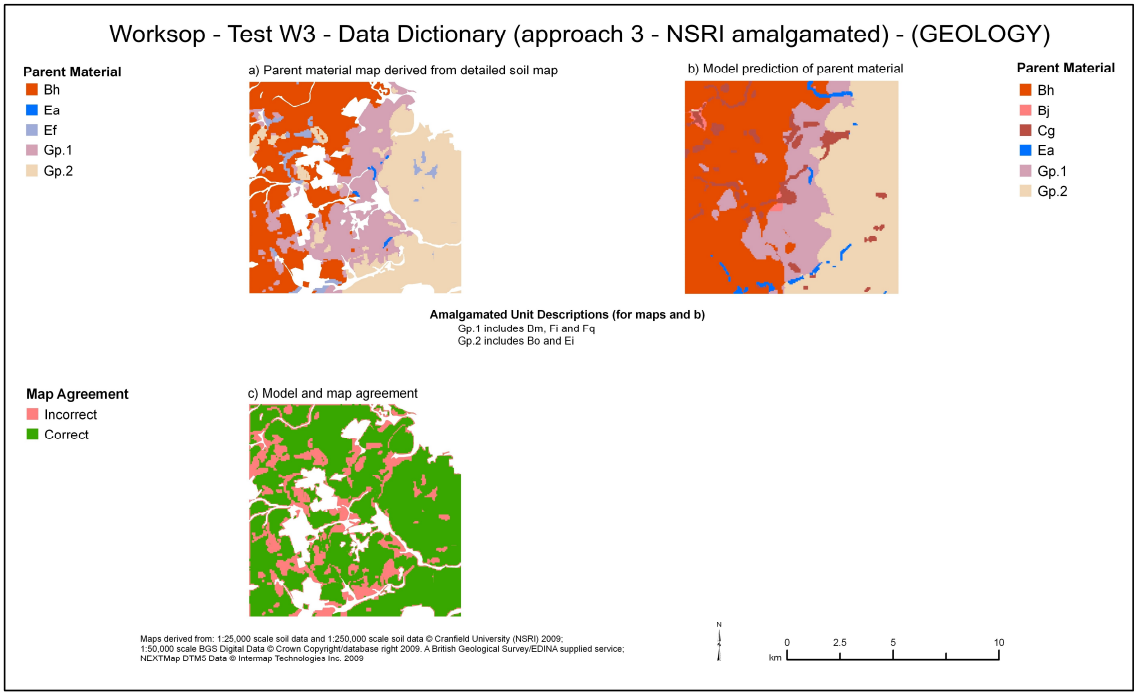


Figure 32 - Test W3 maps (Approach 3)
Input: GEOLOGY (surface); Classification: Amalgamated NSRI PM_LITH; $\Psi_3 = 0.28$; $\theta_1 = 0.40$ $C_e = 3$
A larger version is available in Appendix 4. NSRI PM_LITH codes are described in Appendix 2.

5.6.4 Comparing the bedrock and surface geology inputs

It was hypothesised that the surface geology would outperform the bedrock-only input, yet using the detailed classifications in Approach 1 there was negligible difference between bedrock and surface geology inputs in Worksop and Needwood Forest. In the Yeovil area, a marginally better result was obtained using the surface geology (Test Y1 ($\Psi_3 = 0.15$) versus Y4 ($\Psi_3 = 0.04$), see Figure 33).

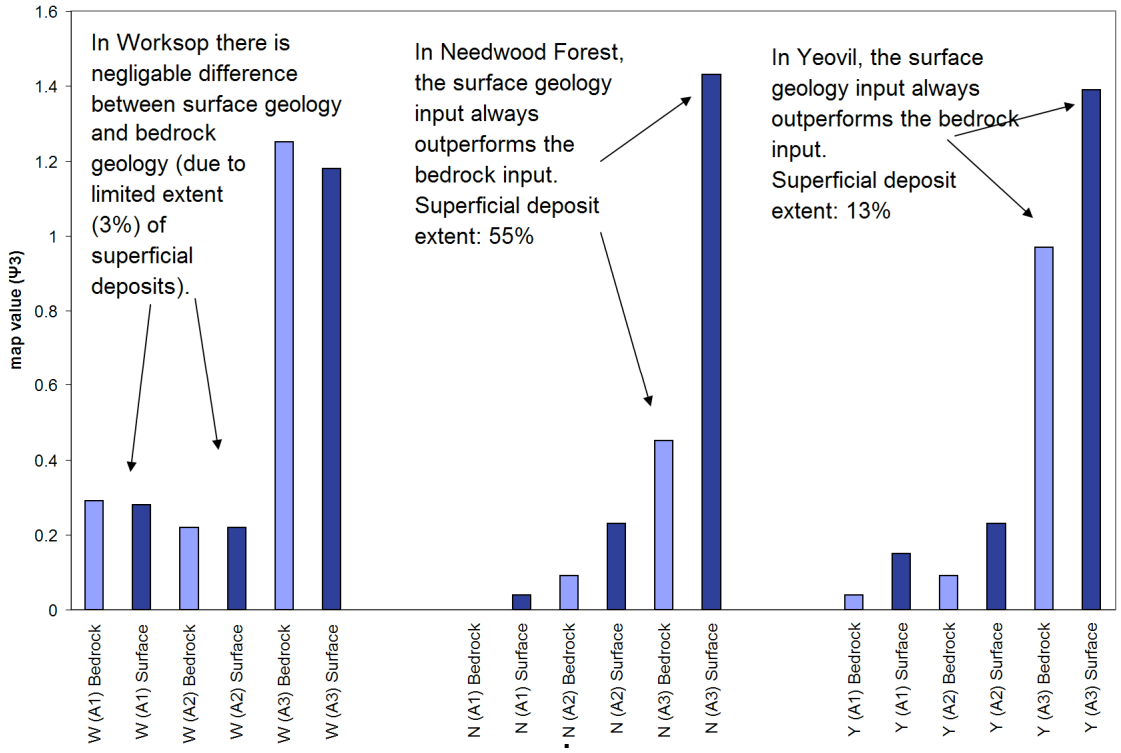


Figure 33 – Comparing the predictive success of bedrock and surface geology inputs
Note: W: Worksop; N: Needwood Forest; Y: Yeovil. A1, A2, A3: Approaches 1 to 3. Tests were performed using the NSRI PM_LITH classification.

With the simplified classification of Approach 2, the surface geology layer produces more valuable maps than just the bedrock input in Needwood and Yeovil. These are the study areas with extensive superficial deposits. But it is when guided amalgamation (Approach 3) is used, that the advantages of the superficial layers become much more apparent. Here, surface geology maps achieve noticeably higher Ψ_3 map values than their bedrock-only comparisons (Figure 33). Additionally, there are consistently more

effective classes (C_e) when using the surface geology rather than the bedrock geology input (cf. Tests Y3 (Figure 34) and Y6 or Y12 and Y18, Table 21).

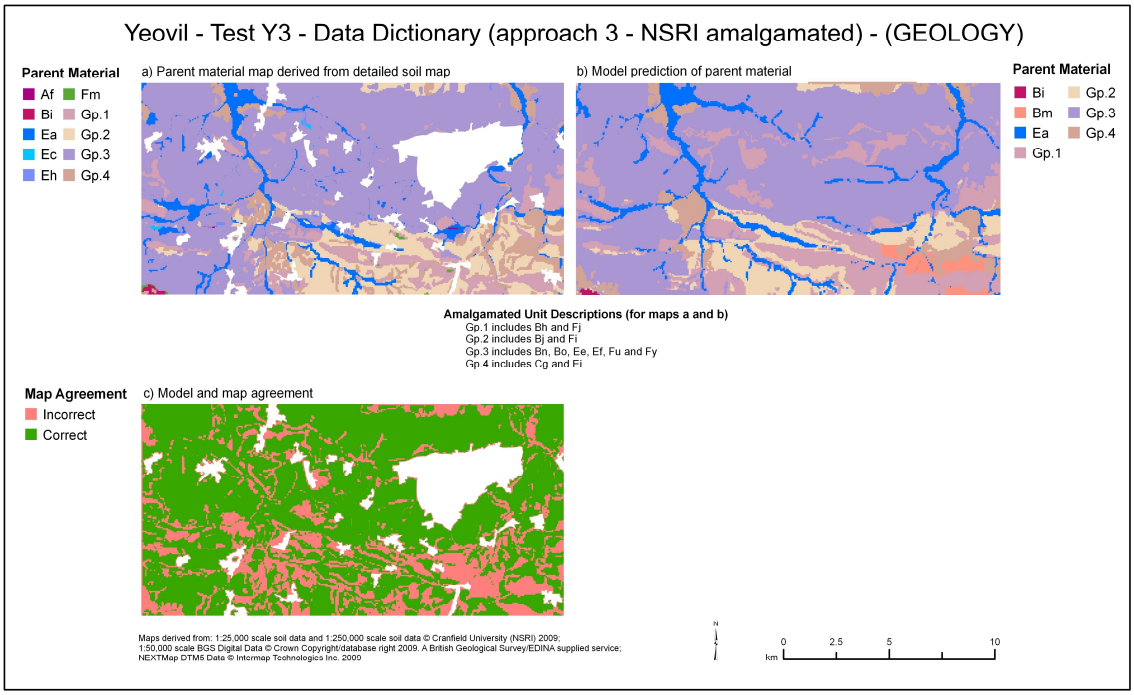


Figure 34 - Test Y3 maps (Approach 3)
Input: GEOLOGY (surface); Classification: Amalgamated NSRI PM_LITH; $\Psi_3 = 1.39$; $\theta_1 = 0.76$ $C_e = 6$
A larger version is available in Appendix 4. NSRI PM_LITH codes are described in Appendix 2.

Removing the superficial geology from the classification can lead to marginally higher levels of agreement between the maps if there is little superficial geology in the area, and the predicting geology map does not depict much of the superficial geology that does exist. This situation occurs in the Worksop study area where the superficial deposits have limited extent (3%), but even here the difference is marginal (Figure 33).

It has been noted that the detail with which superficial deposits are mapped varies across the UK (Palmer et al., 2007). Because of this, and to maintain consistency, one approach could be to derive the parent material map from only the bedrock geology layer. Yet what has been demonstrated here is the considerable benefit that the superficial layer brings to the delineation of the soil parent material. An alternative approach to describing this uncertainty would be to attach a confidence layer to the final map. This will be incorporated into later methodologies.

5.6.5 Assessment of parent material identification

This discussion will highlight the parent material units which tend to be classified correctly and those which are commonly misclassified. In addition, it will provide commentary on possible reasons for the success, or lack thereof.

5.6.5.1 Worksop parent material units

Using the NSRI classification (Tests W1 – W6 (Table 19)) parent material Bh (limestone) consistently performed very well in the Worksop area, achieving class values (ξ) of up to 0.90. This is not reflected in the standard ESB classification (Tests W7 & W13) where there was confusion between dolomite (2120) and limestone (2110). These lithologies are very similar, and indeed, the NSRI parent material classification makes no distinction between them. This wider lithological class accounts for the better agreement. Dolomite is mostly calcium magnesium carbonate ($CaMg(CO_3)_2$), whereas limestone is predominantly calcite, which is calcium carbonate ($CaCO_3$). Under certain conditions, the calcite in limestone can be partially replaced by dolomite forming dolomitic, or magnesian limestone. This has occurred in the Worksop area, where much of the western region is magnesian limestone. These mineralogical misclassifications in the ESB classification are rectified when simplified to group level (Tests W10 and W16) and also through guided amalgamation (Tests W12 and W18).

The best performing units in the standard ESB classification tests were the claystone / mudstone (1310) units which achieved class values (ξ) of 0.65. Interestingly, in the NSRI classification, this lithology only performed well when strongly amalgamated (W3) where 'mudstone, sandstone and slate' (Bm), 'clay or soft mudstone' (Fi) and 'sand or soft sandstone' (Fq) were all combined. This gave a ξ of 0.75 or a weighted class value (ω) of 0.28, when the number of amalgamated parent material classes was taken into account.

There was significant misclassification between predicted Bo (sandstone) and what was mapped as Ei (drift with siliceous stones) in Test W1. The ESB units were also misclassified in Test W8 – 5000 (unconsolidated deposits) versus 1210 (sandstone). These misclassifications can be explained by the nature of the sandstone in the Worksop area. In this region, the Nottingham Castle Sandstone Formation (previously known as the Bunter Pebble Beds) is extensive. This sandstone is characterised by an abundance of rounded, and commonly milky white quartzite pebbles (British Geological Survey, 2009). These pebbles, being chemically and physically robust, survive the erosion of this formation, and form the basis of many locally reworked deposits, as well as being found throughout the country as a result of transport by glaciers and rivers. It is these pebbles which are the basis for the reworked deposits mapped by the soil surveyors as "drift with siliceous stones" (Ei). As can be seen in Figure 31, there is very good spatial agreement between the Ei and Bo units, which leads to a high class value upon amalgamation. It is probable that the geologist made no differentiation between the consolidated conglomerate and the locally reworked Bunter pebbles, while to the soil surveyor the parent material classification demanded that this be placed in the 'Soils in thick drift' category (E) as these soils formed from a layer of Quaternary deposits at least 80 cm thick (Clayden and Hollis, 1984, p19) .

The central mapping unit in Figure 31 (Test W1) reveals a consistent misclassification of Fi (clay or soft mudstone) as Bm (mudstone and sandstone or slate). There are similarities in the lithological descriptions, with both referring to mudstone. The differences arise in the description of the physical nature of the parent material. Bm is assumed to have a lithoskeletal substrate, while Fi has a soft, pre-Quaternary substrate within 80 cm (Clayden and Hollis, 1984, p21). Fi does not occur on the reference map. While the geology maps provide usable lithological information, they do not appear to provide enough information on the physical nature of the top metre to accurately describe the broad parent material type (as defined in Table 13.)

In summary using this methodology there appear to be three main discernable parent material groups in the Worksop area; (1) limestone, (2) clay / mud / sandstone, and (3) a more pure sandstone.

5.6.5.2 Needwood Forest parent material units

Throughout the NSRI tests, peat (Aa) performed moderately well, with a ξ of 0.69. Given that this is a unit of limited extent, making up less than 1% on the soil map, such agreement is notable. There was minor confusion in the ESB classification between peat descriptions (test N8, Table 6). River alluvium (Ea), which makes up 5% of the soil map also performs moderately well, with a ξ of 0.66 (test N1). However, both peat and alluvium are quite distinctive. Peat is an organic parent material, while alluvium is constrained tightly by geography. This distinction may explain the better results obtained for these units compared to the more extensive parent material types in the area. This pattern will be examined in more detail at a later stage.

'Drift with siliceous stones' (Ei) makes up 52% of the Needwood Forest area, and was mostly misclassified as gravelly 'sandstones, siltstones, mudstones or slate' (Cg) or as lithoskeletal 'mudstone, shale or slate' (Bj) (test N1, Figure 24). There is an important distinction in these two misclassifications. It appears that the Cg unit incorrectly describes the correct parent material (Ei), but the Bj unit refers to the bedrock geology which underlies the superficial deposits. Thus, while the Ei / Cg misclassification is terminological, and can be easily rectified, the Ei / Bj misclassification represents a fundamental difference in the mapping of the extent of superficial deposits between the geology and soil maps. Simply put, the reference soil map shows more extensive superficial deposits than the geology map. This is due to a greater emphasis on the parent material of the soil and the larger mapping scale of the 1:25,000 scale reference maps.

The soft, pre-Quaternary 'clay or soft mudstone' (Fi) which, at depth, underlies most of the area makes up about 28% of the soil map. This unit was also mostly misclassified as Bj (mudstone, shale and slate) and Cg (sandstone, siltstone mudstone or slate) in Test N1. In this case, the Fi / Bj misclassifications are terminological and can be overcome, while the Fi / Cg misclassifications represent where the geological and soil maps are at variance. Using the simplified NSRI classification (test N2), a grouping of the lithologically similar units, for example, Fi+Bj and Ei +Cg, led to a moderately good

class value (ξ = 0.66) for 'Clayey Rocks' (Class 11), and yet extensive misclassification remained between 'Quartz and siliceous stones' (Class 5) and 'non calcareous drift and gravel' (Class 23).

Using guided amalgamation (N3), these contentious units were grouped into a large group. This successfully overcame the fundamental differences in the soil and geological maps, and led to significant map agreement for this unit (ξ = 0.98) but at the cost of a distinct and significant reduction in class detail (Figure 35 compared with Figure 24, which shows the results from N1).

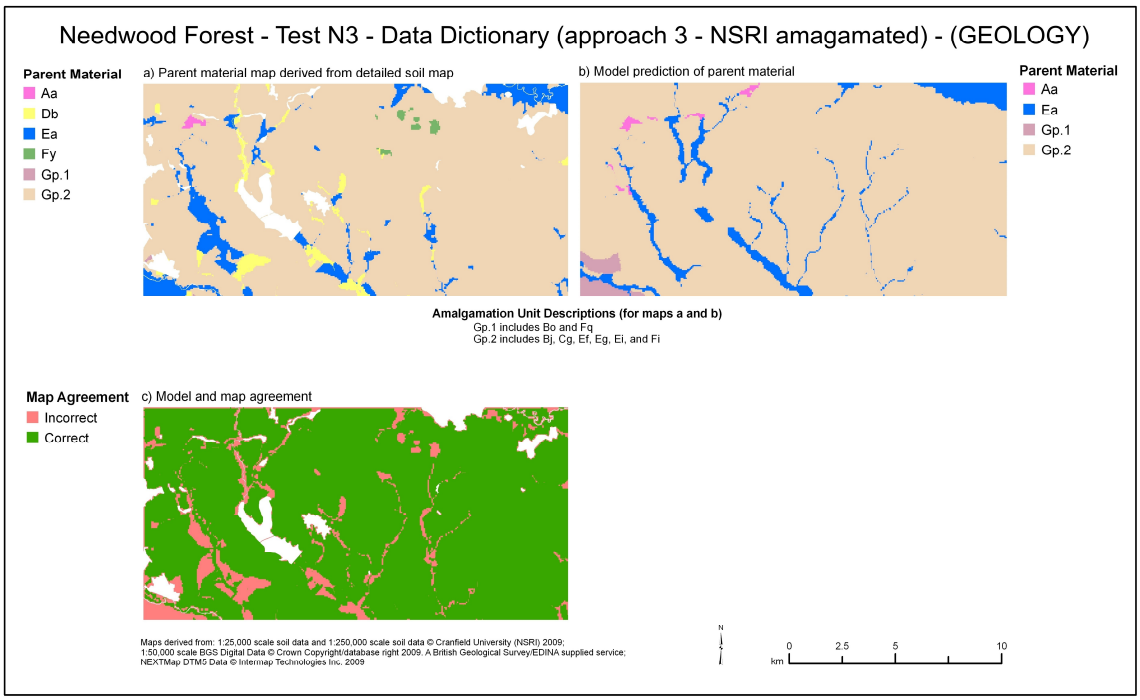


Figure 35 - Test N3 maps (Approach 3)
Input: GEOLOGY (surface); Classification: Amalgamated NSRI PM_LITH; $\Psi_3 = 1.43$; $\theta_1 = 0.95$ $C_e = 4$
A larger version is available in Appendix 4. NSRI PM_LITH codes are described in Appendix 2.

5.6.5.3 Yeovil parent material units

River alluvium (Ea) which makes up about 9% of the soil map, also performs moderately well (ξ = 0.69) in the Yeovil area, with only very minor confusion with sandstone (Bo) (test Y1, Table 7). As sandstone covers 44% of the modelled map, this confusion represents a small relative under-prediction of alluvium by the geological

map. Chalk (Bi) is a distinctive, easily identifiable unit with very limited extent (0.1% of the soil map) and was quite well predicted ($\xi = 0.65$).

Limestone (Bh) covers 8% of the soil map and achieves a ξ of 0.61. There was some confusion with 'clay or soft mudstone' (Fi) and also 'clay with interbedded limestone' (Fj) amongst others. It is likely that the former represent mapping variances, and the latter, terminological confusion. This view is supported by the fact that the clay unit (1310) in the ESB classification are more accurately translated in Test Y8 (Figure 26).

According to the soil map, soft, pre-Quaternary 'loam and soft siltstone' (Fu) and 'soft shale or siltstone' (Fy) are extensive in the Yeovil area, covering 23% and 29% of the area, respectively. However, these units were completely misclassified. What is mapped as 'loam and soft siltstone' (Fu) on the soil map is shown mostly as sandstone (Bo) on the modelled map (test Y1, Figure 23). The predicting geological unit is the Bridport Sand Formation (BDS-SDST), which is described as being predominantly Jurassic sandstone (British Geological Survey, 2009). In this case there is both confusion in the identified lithology (sandstone instead of siltstone) and the nature of the parent material (lithoskeletal substrates instead of soft, pre-Quaternary substrates).

To understand this confusion, additional Lexicon fields were requested and received from BGS. When the supplementary fields of the BGS Lexicon were queried in more depth, the following lithological description, which is much more similar to the parent material description as shown on the soil map, is revealed:

"Grey, weathering yellow or brown, micaceous silt, very fine sand and fine sand, locally with calcite-cemented sandstone beds and lenses, variably sandy clay/mudstone at base, including Downcliff Clay of type area" (British Geological Survey, 2009)

Because this fine grained material was ignored in the simple description of the formation, the unit was misclassified in both the NSRI and ESB classifications. For example, in the ESB subtype classification (Test Y8) sandstone (1210) was predicted to

cover 44% of the area (it should cover 23%, according to the soil map). Much of this area should have been mapped as siltstone (1320).

About 60% of the 'soft shale or siltstone' (Fy) which occurs on the soil map, was modelled incorrectly as sandstone (Bo) by the predicting Dyrham Formation (DYD-SDST). Much of the remainder was misclassified as 'siltstone and sandstone' (Bn). While the Bn classification is closer lithologically to the reference parent material (Fy), differences in the physical nature of the material remain. The Dyrham Formation is also described as Jurassic sandstone, but once more, the additional fields of the BGS Lexicon reveal a more detailed description of the lithology:

"Pale to dark grey and greenish grey, silty and sandy mudstone, with interbeds of silt or very fine sand (locally muddy or silty), weathering yellow. Variably micaceous. Impersistent beds or doggers of ferruginous limestone (some ooidal) and sandstone, which tend to occur at the top of sedimentary cycles. Sporadic large cementstone nodules." (British Geological Survey, 2009)

Once more, this silty and muddy formation is misleadingly summarised as sandstone, when in fact, the sandstone only accounts for a minority of the volume of this unit.

'Clay or soft mudstone' (Fi) was commonly misclassified as 'mudstone, shale or slate' (Bj) in test Y1. This is not a lithological misclassification, but represents a lack of knowledge of the physical structure of the parent material. The soil map recorded this as a soil forming in thin drift (F), and the prediction was a lithoskeletal soil (B) (Table 13). This misclassification was corrected in Approach 2 when using the simplified NSRI classification (Test Y2, Class 11).

5.6.6 Evaluation of the data dictionary methodology

The most valuable parent material maps were produced in all cases using the PM_LITH national parent material classification and guided amalgamation of classes (Approach 3). The analysis of the European ESB parent material maps did involve two translations; NSRI to ESB and GEOLOGY to ESB, which may give rise to some additional misclassifications. Nevertheless, the ESB classification does have some advantages over the NSRI national classification. It is an international system which can be used across all of Europe, is strongly lithological and allows sub-dominance. This allows easier links from geology to be made. However these advantages do not necessarily make it a better parent material classification.

The possibility for European harmonisation comes at a price of a reduction in accuracy. For those working at international scale of 1:1,000,000, it may be that the ESB classification is ideal. Indeed, it was for maps at these more general scales for which this classification was created (FAO, 1995). However, this study has provided a unique quantification of the reduction in detail brought about by the harmonisation of a national to an international classification. While it is true that the NSRI classification has the 'home advantage' in terms of scale and classification, there are clear differences in the level of detail recorded in these maps (as shown by the Ψ_3 values) even with similar overall accuracies (θ_1). As well as producing maps with consistently lower Ψ_3 values (Table 22), the lithological ESB classification also loses a great level of detail about the physical nature, consolidation or cohesiveness of the parent material which is intrinsic in the NSRI classification of soil parent materials and soil series. Similar issues are likely in other national classifications. Therefore, for projects predicting parent material or soil at detailed scales in England and Wales, it is usually preferable to use the national classification. Should the international classification be required, this may also be supplied as supplementary attribution or by means of an additional lookup table.

Table 22 - A quantitative comparison of the loss of detail which is brought about by the conversion from a national parent material system to an international system.
Note: Comparing the map value (ψ_3) and overall accuracy (θ_1) of amalgamated tests using surface geology as a predictor of parent material. (cf. Tests, W3, N3, Y3 & W12, N12 and Y12)

	NSRI		ESB	
Study Area	ψ_3	θ_1	ψ_3	θ_1
Worksop	1.18	0.83	0.78	0.80
Needwood Forest	1.43	0.95	0.42	0.90
Yeovil	1.39	0.76	1.03	0.69

For projects in other European countries, a judgement will be required as to whether or not the ESB classification will provide enough detail about the parent material to meet particular project objectives. If these objectives rely on a purely lithological description of the subsurface, the ESB classification is likely to suffice. If however, there is a requirement to characterise the near surface hydrology, or define soil classes on the basis of the parent material, as is often required in digital soil mapping exercises, it may well be that the ESB classification is not ideal. It may be preferable to use a national parent material classification with subsequent translation to international classifications as required.

In the Needwood Forest and Yeovil areas, the most valuable maps were produced using the surface geology layer, representing the strong effect that superficial deposits have on the soil parent material in these regions. In the Worksop area, only 3% of the area on the geological map has superficial deposits, and in this region, the bedrock input marginally outperformed the surface geology layer (Test W6 ($\Psi_3 = 1.25$) versus W3 ($\Psi_3 = 1.18$)).

These results indicate that the overall accuracy of these maps are approaching a usable level ($\theta_1 > 0.80$), while maintaining a certain level of parent material class detail. However, the number of effective parent material classes is still low, approximately 4 in all areas, when in reality there are up to 17 parent material classes present.

Particularly distinctive parent material types, such as alluvium, peat or chalk tend to achieve high class values, even when their physical extents are small. These types of materials are easily identifiable to both soil surveyors and geologists, and there is little ambiguity in their definitions or physical extents.

Other parent materials, such as sandstone, mudstone and siltstone are more open to confusion between the reference soil map and the modelled parent material. These can arise from the natural complexity of the geological succession, and the interdigitation and gradation between fine and coarse textured sedimentary rocks. Nevertheless, often such class definitions can be overcome with some expert judgement or class amalgamation at the expense of loss of class detail.

There remain fundamental problems with the inconsistent mapping of drift deposits between soil surveyors and geologist. These lead to problematic units where bedrock and superficial derived parent materials are combined, and such classes are broader than is ideal, for example, see Test N3 (Figure 35) where unit Gp. 2 covers the majority of the area. The lack of detail on the geological map about the physical nature of the top 100 cm leads to many of the classes being lithologically similar, yet assigned incorrect broad parent material classes as described in Table 13.

A higher number of effective classes, and particularly classes with even higher weighted class value (ω), brought about by more tightly-defined membership are desired. It is recognised that, given only the single geological dataset as a predictor of the parent material, a perfect agreement is impossible as there are differences in the map scale and detail of the linework. Furthermore, one geological unit may give rise to multiple soil parent material units. Such one-to-many relationships are more complex than the one-to-one translations examined in this methodology, where one geological unit may only give rise to one parent material.

There are weaknesses in this approach resulting from this simplistic one-to-one translation, and a lack of prior knowledge of which parent materials are likely to be found in the study areas. Without this knowledge, large proportions of the maps were misclassified from the outset as parent material classes which are not found in the area. It is likely that additional sources of expert knowledge are available which may enable better initial translation from geology to parent material. The use of such sources should be considered. The addition of further environmental layers, such as geophysical remote sensing data, regional soil maps, aerial photography or digital elevation models might fill in the detail which is missing from the single, geological input.

Key points:

- The national NSRI classification produces more valuable maps than the international ESB classification
- The surface geology map tends to produce more valuable maps than the bedrock only map, particularly in the Needwood Forest and Yeovil areas where drift deposits are extensive.
- The initial classifications of parent material units from geological units are commonly incorrect due to lack of descriptive detail on geological maps.
- Guided amalgamation of misclassified units outperforms lithological simplification of the parent material classifications

5.7 Recommendations

Resulting from this methodology, the following recommendations are made:

- Attempts should be made to better define the relationships between the geological units and the soil parent material units within the study area using additional sources of information.
- Environmental datasets in addition to geology which may also influence or reflect soil parent material should be included in the prediction of parent material.
- Ways of better defining these environmental relationships should be explored, for example:
    - Does local expert knowledge captured in published soil records, auger bore records, soil survey field notebooks, databases or on the Internet help define the relationship between soil parent material and geology in a local context?
    - Can machine-learning characterise the relationships between soil parent material and geology, as well as other environmental layers which might be used as correlatives for soil parent materials? These correlatives might include digital terrain model derivatives or regional scale soil maps
- A probability model should be created to combine a variety of potential parent material covariates to predict the likelihood of any given parent material.
- The use of the NSRI parent material classification is recommended for environmental modelling projects in England and Wales.
- The use of surface geology datasets are recommended over the use of bedrock only datasets, particularly in regions where there are extensive superficial deposits.
- It is recommended that classifications are simplified on the basis of reference area sampling and the use of guided class amalgamation.
- It is recommended that an assessment of the likelihood of accurate prediction be made available for each parent material class so that knowledge of errors can be propagated.

6 EXPERT KNOWLEDGE METHODOLOGY

The data dictionary methodology demonstrated that surface geology has the potential to be a good predictor of soil parent material, particularly when misclassified units are amalgamated. Nevertheless initial predictions of parent material were often erroneous. To improve these predictions, the expert knowledge methodology uses expert knowledge captured in published literature to better define the relationships between parent material, geology, and additional environmental covariates to further enhance prediction of parent material. This methodology considers the probability of each parent material class, on the basis of three evidence layers; geology, slope and the National Soil Map.

6.1 Introduction to the expert knowledge methodology

This methodology tests the value of extracting expert knowledge about the relationships between environmental covariates and parent material, and using this to guide predictive models of parent material. Probability model inputs were built using expert knowledge extracted from books, supplemented with additional information from national databases. The aim of this method was to discover how much information could be gleaned from published sources and how useful this knowledge would be in predicting the soil parent material, given a small range of environmental datasets.

Where possible, qualitative information from the published literature was quantified, and probabilities of occurrence were derived for each parent material given the environmental evidence (e.g. a 'gentle' slope or a certain geological unit). Probabilities were combined for all data layers using a modified and corrected probability model based on the Expector method (Corner et al., 2002; Farewell and Farewell, 2010).

6.2 The use of expert knowledge in environmental models

Experienced practitioners, in any area, tend to develop an intrinsic understanding of their subject. This is certainly the case in mapping sciences, such as soil and geological survey, where field surveyors develop a mental model of the landscape (Bui, 2004) and the ongoing processes within it. A wealth of knowledge, not necessarily available through final map products is contained within the field surveyors' mind, and to a lesser extent, in their notebooks, published records and derived information sources. If extracted and formalised, this information has the potential to be of assistance in generating parent material maps. However, it has been noted that soil surveyors can be poor at stating their mental models explicitly (Lagacherie et al., 1995).

In the data dictionary method, the translation from geology to parent material was based entirely on the information contained on the maps, which led to a consistent misclassification of units. In this methodology, the use of expert knowledge and additional environmental layers will be investigated to determine if parent material maps of higher value (ψ_3) can be created.

6.2.1 Techniques of acquiring expert knowledge

There are a number of possible techniques of extracting and formalising expert knowledge. In this study knowledge of the relationships between environmental covariates and parent material is of interest. Knowledge can be formalised through interviews, both structured and informal, or extracted from published literature containing block diagrams and prose, databases and the Internet.

The interview is the most commonly used method of extracting and formalising expert knowledge, as this approach allows the interviewer to ask, and receive answers to specific questions. Zhu et al. (1996) used a structured interview approach to acquire expert knowledge about the distribution of four soil series within a landscape. They found this technique to be effective at revealing knowledge, but that it was very time

consuming. On the basis of the interviews they created curves showing the relationship between a particular soil series and a range of environmental covariates, such as elevation. This approach relies on the availability and willingness of the original surveyors to engage with this interview process.

Interviews can help formalise environmental relationships, but rely on the presence and participation of the expert. Most soil surveyors in England have now retired, but there remains a wealth of information about soils captured in published books. Extracting relationships from published sources relies on the author consistently, adequately and accurately describing the relationships of interest.

Expert knowledge captured by expert input into a GIS has been used in habitat and species distribution modelling. Yamada et al. (2003) compared two approaches of extracting expert knowledge from nine park rangers regarding the distribution of samba in Victoria Park, Australia. They found that a semi-structured interview approach outperformed a quantitative approach of data input by the same rangers using a GIS and also that more consistent results were reached between rangers. They warn against reliance on only one source of expert knowledge. To predict suitable habitat for an endangered species, Smith et al. (2007) used a combination of expert knowledge and limited empirical field data with a Bayesian belief network. They used published literature to develop conceptual models prior to the development of their habitat models. They discovered that available expert knowledge was a useful surrogate for empirical data, and comment that Bayesian belief networks offer a flexible basis for combining expert knowledge with empirical data.

McKenzie and Ryan (1999) note that the intuitive expert knowledge gained by a soil surveyor traversing a landscape greatly increases the predictive accuracy of a conventional soil survey, even in an unfamiliar areas, and yet such mental models are difficult to capture for explicit models.

Models which involve expert knowledge tend to incorporate Bayesian logic, where the probability of a certain class is calculated on the basis of input data, or fuzzy logic (Zhu

et al., 1996) where the relative membership of classes in an area or wider class can be considered. This is a significant progression, in terms of the possible creation of a parent material map, from the data dictionary methodology, as such methods introduce probabilities or mixed classes to the resulting model. These may be appropriate given the complex nature of interactions between geology and soil parent material seen in the data dictionary methodology. McBratney et al. (2003) suggest that quantitative modelling of soils classes and attributes by expert knowledge is an area in which more research is required.

Hansen et al. (2009) suggest that published papers characterising soil profiles in their Ugandan study area could, with careful reading, yield rules for expert soil mapping techniques. They used the published expert knowledge of soil landscapes to subdivide the area into four landscape classes. However very few, if any, studies have used expert knowledge contained within published literature to populate or build predictive models of parent material. As parent material is a key input into a range of environmental applications, this is an area which could benefit from further investigation.

A number of approaches from the related disciplines of mineral exploration and digital soil mapping have been developed which integrate expert knowledge with sampled data. PROSPECTOR (Duda et al., 1978; Katz, 1991) was an early expert knowledge system which used surveyors expert knowledge combined with probability modelling to predict patterns in the natural environment. The Expector method and software (Corner et al., 2002) was a development on the PROSPECTOR approach. Expector, based on Bayesian Theorem, uses conditional probabilities to assess the relationships between evidence layers and hypotheses (such as soil classes). Expert knowledge can be used to alter the input conditional probability tables and the model provides probabilities for each hypothesis based on classes of the evidence layers. Additionally, it provides a mechanism for dealing with the uncertainty associated with elements in the input datasets.

Expector was judged to offer many of the features which would be useful in integrating expert knowledge gleaned from published literature with predicting spatial data,

although there were errors in this model. These were corrected in Farewell and Farewell (2010) and this corrected probability model was used in this research.

Wielemaker et al. (2001) provide methodologies for formalising the landscape knowledge of soil surveyors by firstly placing terrain objects in a nested hierarchical structure, and secondly applying formalised knowledge rules to these objects in a GIS. This multi-scale subdivision of the landscape is not formally addressed in Expector, although technically evidence layers with difference scales could be combined.

The majority of predictive mapping exercises using expert knowledge have relied on input from the actual expert, and this predominantly through a structured interview process. Because of the curtailment of extensive field survey programmes conducted by the Soil Survey of England and Wales in 1987, most field surveyors with a detailed understanding of the relationships between parent material and the landscape have retired and there have been few replacements. There is, therefore a need to assess to what extent expert knowledge on the relationships between parent material and environmental correlatives can be extracted from published literature, and how this can be employed in soil parent material models.

6.3 Assumptions

The following assumptions were made for the expert knowledge methodology:

- That the 1:25,000 detailed soil maps accurately record the true distribution of soil type and soil parent material.
- That the soil parent material is related to the mapped superficial and bedrock geology, slope and national soil map.
- That the distribution of the parent materials within the study area is unknown for the purposes of modelling.
- That the approximate extent of the parent materials within the study area is known.

6.4 Expert knowledge methodology overview

The expert knowledge methodology was comprised of a number of stages. A suitable modelling approach was chosen to integrate expert knowledge and sampled data. Sources of expert knowledge describing the relationship of environmental covariates with parent material were identified and assessed. Three environmental covariates were identified which could provide a spatial framework for the identified expert knowledge. These were the surface geology dataset used in the first methodology (GEOLOGY), the National Soil Map (SOIL) and a slope class dataset, derived from digital terrain models (SLOPE). Following identification of these layers, the knowledge was extracted, updated and harmonised. As much of the knowledge was qualitative, this needed to be quantified for input into the models. Key aspects of this method of extracting, formalising and using expert knowledge to predict parent material are shown in Figure 36.

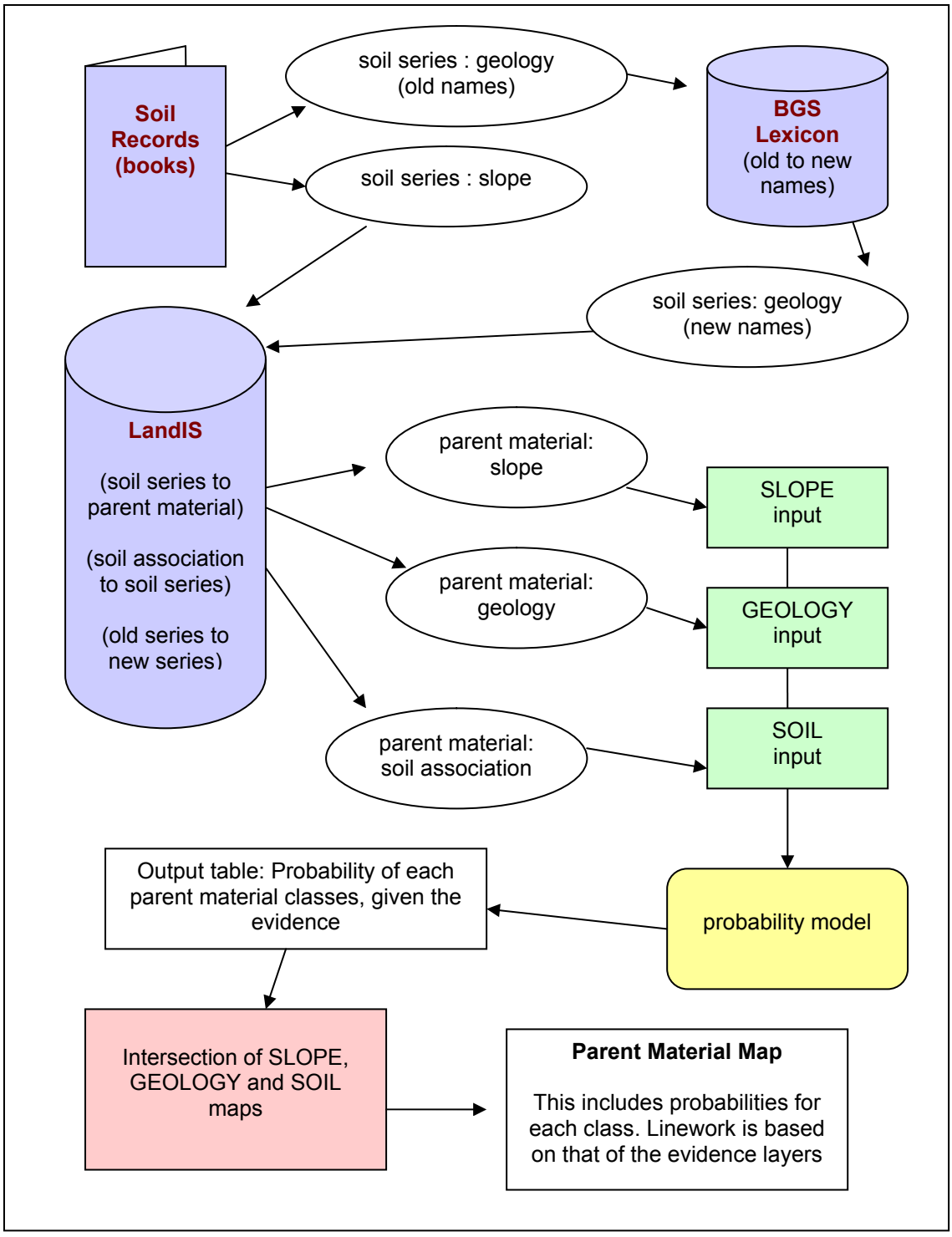


Figure 36 - Production of parent material map from expert knowledge

Note: blue represents sources of expert knowledge. White ovals represent expert knowledge of a relationship. Green squares represent quantified model inputs based on the expert knowledge, into the (yellow) probability model. The output from this model is joined to the (pink) base map which is an intersection of the three evidence layers (GEOLOGY, SLOPE, SOIL). This join produces the parent material map.

Once the evidence data layers, (GEOLOGY, SLOPE, SOIL) and their relationships to parent material were prepared, a number of models with different evidence layers and weightings were run. The results were mapped and assessed in an identical manner to the first, data dictionary methodology, with the addition of model confidence maps to the mapped outputs. This addition of information regarding the probability or certainty of the prediction is useful as it allows knowledge of errors to be easily propagated. Classes with consistent misclassification were amalgamated for a second suite of tests to improve map value.

As this methodology exclusively used the expert knowledge extracted from literature and databases to derive the probability model inputs, the use of spatial data analysis was avoided. Therefore, neither the extent of evidence layer (SOIL, GEOLOGY, SLOPE) class nor the areas of joint occurrence of the different map units between data layers were used in the creation of the model inputs. These relationships are, however, used in the third, data mining methodology.

6.5 Identifying sources of expert knowledge

The first stage in this methodology involved the identification of suitable sources of expert knowledge. A range of sources containing expert knowledge of the environmental relationships within the study areas were identified. These are described briefly in Table 23 (soils information) and Table 24 (geological information). A range of soil literature was available describing local soil distribution with reference to the underlying geology on a detailed (Soil Records) and regional (Soil Bulletins) scale. This information was supplemented by non-spatial information on the mapping techniques, and class definitions (Soil Monographs and Handbook). Some key aspects of this information, such as the relationship between soil and parent material has already been captured in the LandIS database (Keay et al., 2009).

An assessment of potential value of the knowledge was determined on the basis of the depth, breadth and consistency of the expert knowledge, and the ease with which it

could be extracted and applied to a modelling context. Easily categorised or quantifiable information was also deemed to be more valuable than qualitative descriptions, as these are easier to transform into quantitative model inputs. Only those sources with a moderate or high value were used in the creation of models.

Table 23 - Sources of soils expert knowledge

Source of Knowledge	Relationships Described	Comments	Value
Soil Survey Records (books)	Soil series to slope, geology, drainage and a range of other attributes	Soil series based, so a translation of soil series to parent material is required. This can be found in LandIS.	High
Soil Survey Regional Bulletins (books)	Local composition of soil associations by soil series, as well as descriptions of the landscape, landuse, drainage etc.	Soil Association based landscape descriptions	Low
Soil Survey Technical Monographs (books)	Only general comments	Descriptions of parent materials. Little about relationship to other variables.	Low
Soil Survey Field Handbook (books)	Slope % to slope classes used in the published literature.	Descriptions of slope classes	Moderate
LandIS relational database (soil database)	1) Soil Association Composition 2) National parent material - slope distribution 3) Soil series to parent material conversion	National soil database – extensive information but mostly national averages	Moderate
Soil Survey Archive	Unknown	Hard to access	Unknown
Soil Survey Field Sheets	Soil classes and topography	Hard to interpret, key required. Limited access.	Low

Table 24 - Sources of geological expert knowledge

Source of Knowledge	Relationships Described	Comments	Value
Geological Survey records (books)	Lithologies, colours, ages, boreholes	Little information on overlying soil	Low
Detailed BGS Lexicon (geological database)	Old and new geological names, detailed lithologies of geological units	Very detailed but complex structure	Moderate
BGS archive	Unknown	Restricted access	Unknown
Geological Survey Field sheets	Unknown	Restricted access	Unknown
Internet	Local stratigraphy, older geological names	Variable reliability, helps in the sequencing of local geological beds from diagrams	Moderate

The source of expert knowledge which was found to have the most extensive information on the relationship between environmental covariates and parent material was the Soil Survey Record for each of the 1:25,000 detailed soil maps (Jones, 1983; Reeve, 1976; Colborne and Staines, 1987). These Records aim to describe representative soil landscapes across England and Wales. The Records are not consistent in the detail of descriptions of the various soils and landscapes, but do contain a wealth of information. An example of the type of information which is found in these records is presented in Figure 8 on page 26. This prose was extracted from the text and formalised into a more structured dataset (Digital Appendix 1).

While the Soil Survey Records describe the relationship between underlying geological units and the soil, the geological records tend not to describe the soil parent material in any detail. Thus, these geological books are of little value in adding further information to that which can be obtained from the detailed British Geological Survey (BGS) Lexicon (British Geological Survey, 2009).

LandIS (NSRI, 2009; Keay et al., 2009) is the Land Information System for England and Wales. It contains a wide range of soil attribute datasets. Of particular interest here are the translation tables between soil series and parent material classes. Additionally, the memberships of the soil associations (map units of the National Soil Map (SOIL)) are described, providing national averages for each map unit. Caution is warranted, however, as it is not known to what extent the national averages are comparable with local soil parent material membership. There is likely to be significant regional variation in the composition of the map units. The 1:50,000 scale field sheets for the National Soil Map were used extensively in the re-digitisation of the map in 2000. Thus, this more detailed linework has been incorporated into the digital product.

The soil associations as shown on the National Soil Map are described in more detail in the Regional Bulletins. Typically, the soil water regime, cropping patterns and basic descriptions of the soils and underlying geology are provided although the level of detail of the descriptions can be inconsistent. These descriptions are necessarily general and, with regards to the parent material, do not add much additional information to that which can be gained from existing datasets in LandIS.

The soil monographs and handbook (Hodgson, 1997) provide descriptions of methodologies and mapping techniques, as well as descriptions of the NSRI classification system for soils. While key aspects of these books have been formalised in the LandIS database, there is often additional information to be found. Of particular note is Technical Monograph 17 (Clayden and Hollis, 1984) which provides more detailed descriptions of the parent material classification than is contained within LandIS.

The detailed BGS Lexicon is an excellent resource allowing more detailed descriptions of the lithologies of the geological units to be made. Additionally, the Lexicon contains a range of historic names for the geological units. This has been found to be invaluable as many of the names of geological units used by the soil surveyors in their Records are now obsolete.

The BGS have been using their archive and field sheets to achieve better prediction of soil parent material (Lawley and Smith, 2008). In order to avoid duplication of approaches, and as this was a resource which has limited accessibility to this research, the BGS archive was not investigated in this methodology.

At the time of writing, the Soil Survey archive lacked structure and was rather unorganised, physically making the discovery of relevant information very difficult. Recent work, due for completion in 2010 will provide a more structured archive enabling better information retrieval. Once the archive arrives at its new location on the campus of Cranfield University, the potential of archived materials including field notebooks, mapping field sheets and detailed auger bore records should be investigated further.

The Internet offers the opportunity to supplement information missing from databases or published literature, assuming caution is applied with regards to the source of the information. Of particular value are photographs of the landscape and geological outcrops, as well as graphical representations of stratigraphic columns. Such images tend not to be stored within current databases.

6.6 Extracting expert knowledge

Once the sources of expert knowledge had been assessed, the relevant expert knowledge were extracted and compiled. Soil Survey books were scanned and the text extracted using optical character recognition (OCR) software. To maintain flexibility, the information was initially compiled in a Microsoft Excel 2003 spreadsheet. However, due to limited character display in Excel, the knowledge was compiled in the OpenOffice spreadsheet, Calc, which could display numerous paragraphs of text with no difficulties.

The authors of the soil records from the three study areas provided a large variation in the breadth and depth of the recorded information. There were also differences in the

emphasis of the discussion. Nevertheless, broad information classes such as slope, soil water regime, differentiating soil characteristics, underlying geology and lithology were defined in the spreadsheet, and these cells were populated with prose from the books. The collected expert knowledge was queried (sorted, examined and relevant information extracted) to populate a suite of approximately 30 refined fields with more specific information ranging from slope to typical landuses. Three key fields were used in the creation of quantitative models – those pertaining to slope, geology and membership of soil associations. A simplification of one record in these spreadsheets summarising the expert knowledge is displayed in Table 25. The full expert knowledge workbook can be found in Digital Appendix 1.

The LandIS database (NSRI, 2009; Keay et al., 2009) was used to provide descriptions of the parent materials and a range of other soil characteristics for each soil series described in the study areas. This information was added to the spreadsheet holding the knowledge extracted from the published literature.

National compositions of the soil series within the soil associations were also derived from LandIS. Soil associations are the map units for the digital National Soil Map (SOIL) (NSRI, 2008a) which has been used in this methodology to provide information on the regional soil variation. When deriving the input tables for the National Map Units, soil series which, according to SOIL should exist in the area, but do not, were excluded on the basis of expert knowledge from the Soil Records.

Finally, LandIS also provided a percentage breakdown of the recorded slopes for each parent material from approximately 6000 sites on the systematic National Soil Inventory dataset (NSRI, 2008b). These sites are spaced at 5 km spacing and are statistically representative of the soils across England and Wales (Bellamy et al., 2005).

Table 25 - Extract from Expert Knowledge Spreadsheet

Note: This table represents a small part of the full dataset, which is presented in Digital Appendix 1. The information shown below is extracted from two sources: the LandIS database (L) and the Soil Record (SR). From the information collected a prediction of the likely geological units is provided (row 10), along with a measure of confidence (row 11).

1) Rationalised NAME (L)	ABERFORD
2) Sheet	Worksop
3) Series Defined Parent Material (L)	(Bh, B2) limestone
4) PARENT_DESC (L)	Soils over lithoskeletal substrate
5) Geological Unit (SR)	Lower and upper Magnesian Limestone (dolomitic)
6) Lithology / Parent material (SR)	Limestone dipslope (The soil is on both Lower and Upper Magnesian Limestone in the Permian succession. These are dolomitic limestone and dolomites separated by 32 thin mudstone bands in their lower part. Analysis of the heavy mineral suite of the soils (Crampton 1959) suggests that some foreign (i.e. glacially transported) material is incorporated with that derived from weathering of the Magnesian Limestone in place. As the soil profile increases in depth, so the incorporation of foreign material increases, some being far-travelled. Nevertheless, the mineral assemblage of these soils is still very like that of the Magnesian Limestone underneath)
7) Geology	(Bh, B2) Limestone dipslope - Upper and lower Magnesian Limestone (Permian)
8) Slope / landscape	Slopes and dry valleys of the limestone dipslope
9) Parent Material Series	(Bh, B2) Aberford
10) DiGMap Units Best Guess	BTH-DOLM (50%) / CDF-DOLO (50%)
11) How certain?	80%
12) Parent Material Slope / landscape	(Bh, B2) Limestone Dipslope; slopes and dry valleys and depressions of the limestone dipslope
13) Slope / Landscape (SR)	Limestone Dipslope; slopes and dry valleys of the limestone dipslope (with Whitwell (Ipplepen) series)
14) SYMBOL Rationalised (L)	aF
15) Series definition (L)	medium loamy material over lithoskeletal limestone
16) Subgroup trans (L)	typical brown calcareous earths

6.7 Assessing, updating and harmonising the extracted expert knowledge

Once the expert knowledge was compiled in the spreadsheet, the consistency and usefulness of the information as a predictor of parent material was assessed. Much of the information was difficult to quantify or was inconsistent in its recorded level of detail. As such, this information was of little value in the creation of quantitative inputs for the probability models.

For the expert knowledge extracted from the Soil Records, the knowledge was written on a soil series basis. Thus, in the initial expert knowledge table, this information is also collated on a soil series basis. As many soil series can share the same parent material, this information was summarised for each parent material at a later stage.

6.7.1.1 Obsolete and complex nomenclature issues

Due to the age of the literature (Jones, 1983; Reeve, 1976; Colborne and Staines, 1987), many of the descriptions use soil series names and geological names which are obsolete. This made linking between information sources problematic. Therefore, the LandIS and BGS Lexicon national databases were queried to translate the historic names to the modern ones. It was not uncommon for the translations to be complex. These could be caused by a change in the groupings of geological units or many-to-one relationships. Nevertheless, this stage was vital as it enabled the expert knowledge derived from the Soil Records to be spatially joined to the modern geological mapping.

Three additional geological correlatives spreadsheets were established to correlate the geological unit names with those used in the descriptions of the parent material and underlying geology by the soil surveyors (Digital Appendix 1). The geological correlatives spreadsheets were based in part on the more detailed BGS Lexicon (British Geological Survey, 2009). This included information on the previous names of units which had been grouped together to form the new unit. For example, the Soil Records

(Jones, 1983; Reeve, 1976; Colborne and Staines, 1987) make frequent reference to historic units such as the Tea Green Marl. This is only one of more than 5 historic names for what is now called the Blue Anchor Formation – a predominantly mudstone formation.

The hierarchical geological classification employed by the BGS can make reference to the units confusing. For example, the Blue Anchor Formation (BAN-MDST) is a member of the Mercia Mudstone Group, which is, in turn, a member of the New Red Sandstone Supergroup. Confusion may arise when reference is made to the Mercia Mudstone Group, as this is also the name of the mapped Lexicon unit (MMG-MDST), previously known as the Keuper Marl. This too is a member of the New Red Sandstone Supergroup.

Due to complications such as these, diagrams of local stratigraphy obtained from the Internet (West, 2007) were compared with tables of the local geologies created from the BGS Lexicon, sorted by the age of the units. A rough estimate of the confidence of the translation from the modern geological units to those terms used in the Soil Records, was applied in each case. This level of confidence was also stored in this geological correlative spreadsheet (Digital Appendix 1).

6.7.1.2 Missing information

Occasionally, there would be missing information for a small number of soil types or geological units. In these instances, other information sources, such as databases or the Internet were queried to provide the missing information. Where the missing information could not be found from these sources, informal interviews with former field soil surveyors (Jones, 2006) provided the expert knowledge. This type of supplementary information accounted for less than 1% of the summarised data.

6.8 Quantifying the qualitative expert knowledge

After compiling and refining the expert knowledge, only three main classes of knowledge provided enough information to enable the creation of pseudo-quantitative model inputs. These were the relationships between parent material and surface geology, regional soil associations, and slope. There was not enough expert knowledge for other potential classes, such as parent material class elevation ranges, to create the model inputs. In this context, the term pseudo-quantitative has been used to describe the probabilities assigned to relationships based on qualitative data. An entirely quantitative model based on measured properties and relationships might be preferable, but such quantitative data cannot be sourced from expert knowledge.

6.8.1 Quantifying slope datasets

The process of quantifying the mostly qualitative expert knowledge and creating quantitative probability model inputs involved a number of stages. Soil surveyors use a defined sequence of descriptions of slope, ranging from level to precipitous (Table 5), thus it was possible to quantify the descriptions of relationships between soil series and slopes from the soil records. It was not uncommon for soil surveyors to estimate slope in the field, and this had to be accounted for in the model.

Difficulties in the quantification arose with the distribution of the parent material over the slope classes. For example, if the description stated that a certain soil type was found “predominantly on gently sloping land”, accurate quantification of this text is very difficult. In this situation, it was common to provide a distribution of 80% on the gently sloping land and the remaining 20% distributed amongst the other slope classes, favouring those with more similar slopes.

Occasionally in the soil records, due to the heavy use of prose, the authors have been inconsistent with their usage of defined descriptive terms. One example in the

Needwood Forest study area is in regards to the Crannymoor soil series. In the text, there is no recorded information about the slope for this series, except that,

"Where it (Crannymoor) joins the stony phases of the Newport and Bromsgrove series slopes are steep" (Jones, 1983)

Technically, this description (steep) means the slope must be greater than 15 degrees (Hodgson, 1997). Because this seemed rather different to what was intuitively expected for this series, clarification was sought with the author, Bob Jones, who undertook the soil survey. He gave assurance that the Crannymoor series was in fact not found on slopes which were steep, in the technical sense, but rather, just *steeper* than one might expect for that series. He mentioned that at the junction of the mentioned series, the maximum slope would be in the region of 8 degrees (which technically is only a 'strong' slope - see Table 5). The rest of the Crannymoor series, he said, would be in the region of 3 and 5 degrees.

Issues such as these raise concerns about the trustworthiness of information acquired from sources such as Soil Records. Where the English language is concerned, there appears to be some room for misinterpretation, and this needs to be considered and accounted for in the models.

In the example of the Crannymoor series, to maintain consistency, it was decided to leave the description, as derived from the book, with slopes which could be up to steep. Attempts are made to correct this with the addition of quantitative data in subsequent methodologies. Where no slope information was provided by the Soil Records, common sense was applied. For example, alluvial units, where no information was provided, were defined to form predominantly on level slopes.

In two cases, where no information was provided in the literature, and the likely slopes of a certain parent material were less obvious, the representative National Soil Inventory (NSI) dataset (NSRI, 2008b) was queried to obtain the national distribution of slopes for those parent materials.

6.8.2 NSI slope data alternative approach

Because of the concerns previously identified with the expert knowledge data extraction for the slope parameter, SLOPE model inputs were also derived from the national slope distribution for parent materials available in the National Soil Inventory (NSRI, 2008b). For these inputs, the distribution of parent materials across slope classes from the 5,148 NSI points across England and Wales were analysed. This distribution was used to create the alternative NSI SLOPE model input.

6.8.3 Combining soil series information for parent materials

For this stage, to create a soil parent material map, it was necessary to compile and collate the expert knowledge which was originally attributed to the soil series, for each soil parent material unit. Some difficulty arises when the slope ranges or the underlying geology differ between soil series which have the same soil parent material. The proportion of each geology or slope class must be decided for each parent material unit.

As a result of not using spatial analyses to guide the likelihood of membership in a particular slope or geology class of a parent material, the likelihood had to be estimated. In effect, the area each series occupied, or how commonly it occurred on a slope class or geology class was unknown. Consequently, unless strong expert knowledge was overriding, for a particular parent material class, the likelihood would commonly be equal across the slope or geology classes defined for it.

Simple probability inputs, similar to fuzzy logic functions, were created for each parent material class in each study area. These show the distribution of a particular class across a range of evidence (e.g. SLOPE) classes. An example model input is presented in Table 8 (p69).

6.8.4 Probability model runs

A probability model (see section 4.3, p64) was used to combine the probabilities for each model input related to the three environmental covariates (SLOPE, GEOLOGY, SOIL). Models for all combinations of input layers were run. Slope models were initially run at full weight, but these tests showed SLOPE to exert too strong (and incorrect) an influence on the prediction of parent material. Therefore, when combined with the GEOLOGY and SOIL inputs, half-weighting of SLOPE was subsequently applied. The weights for the evidence layers for the tests are shown in Table 26.

Table 26 - Tests and weightings for the expert knowledge methodology
Note: this table shows the weighting of evidence layers for the models for the unamalgamated tests. The same weighting apply for the amalgamated tests 28-36. 1 indicates full weight. 0.5 indicates half-weighting.

		Test Number								
Evidence Layer	**Code**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
Expert Knowledge Slope	EK SLOPE		1							0.5
NSI Slope	NSI SLOPE	1			0.5	0.5		0.5		
Surface Geology	GEOLOGY			1	1	1	1			1
Soil Association	SOIL					1	1	1	1	

6.8.5 Model Outputs

Given the evidence layers supplied (SLOPE, GEOLOGY, SOIL), the models output tables which describe the probability of membership in each parent material class for that specific combination of evidence layer classes (see example in Table 9, p70). From this table, the most likely parent material was identified for each combination of evidence attributes, along with the associated probability of prediction.

The output file for each model run was joined to a systematic 60 m point sample shapefile containing the attribute data of all the input evidence layers. A 60 m grid was

chosen as this was the maximum point density achievable for later analysis within Excel 2003. Agreement was then calculated between the most likely parent material predicted by the model and the actual parent material, according to the reference map, allowing agreement maps to be made. The probability of the most likely parent material was used to create a model confidence prediction map. The agreement and model confidence maps were used to determine trends in the behaviour of the models.

6.9 Expert knowledge methodology results

The results from the expert knowledge methodology are presented in Table 27 to Table 29.

Table 27 - Results for expert knowledge methodology – Worksop:
Note: For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Worksop Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
NSI SLOPE (full weight)	-0.01	**0.18**	**0.04**	9	3	2	-	-	-	-	1	58%	W19
EK SLOPE (full weight)	-0.06	**0.14**	**0.02**	9	5	1	-	-	-	-	1	13%	W20
GEOLOGY	0.35	**0.48**	**0.57**	9	5	3	3	3	-	-	1	28%	W21
GEOLOGY, NSI SLOPE	0.34	**0.47**	**0.55**	9	7	3	3	3	-	-	1	11%	W22
GEOLOGY, NSI SLOPE, SOIL	0.38	**0.50**	**0.74**	9	8	3	3	3	-	-	1	6%	W23
GEOLOGY, SOIL	0.38	**0.51**	**0.76**	9	7	3	3	3	-	-	1	21%	W24
NSI SLOPE, SOIL	0.51	**0.61**	**1.18**	9	5	4	3	3	1	-	1	13%	W25
SOIL	0.51	**0.62**	**1.19**	9	4	4	3	3	1	-	1	28%	W26
GEOLOGY, EK SLOPE	0.27	**0.39**	**0.37**	9	6	3	3	3	-	-	1	13%	W27
NSI SLOPE (full weight) (A)	0.11	**0.92**	**0.34**	3	2	1	1	1	1	1	7	0%	W28
EK SLOPE (full weight) (A)	0.04	**0.72**	**0.28**	6	2	1	1	1	1	1	4	13%	W29
GEOLOGY (A)	0.76	**0.84**	**1.21**	5	4	3	3	3	2	3	3	2%	W30
GEOLOGY, NSI SLOPE (A)	0.74	**0.83**	**1.16**	5	5	3	3	3	2	3	3	0%	W31
GEOLOGY, NSI SLOPE, SOIL (A)	0.75	**0.84**	**1.32**	6	5	3	3	3	2	2	3	6%	W32
GEOLOGY, SOIL (A)	0.79	**0.87**	**1.32**	5	5	3	3	3	2	2	4	2%	W33
NSI SLOPE, SOIL (A)	0.77	**0.84**	**1.84**	6	4	4	4	4	2	2	3	2%	W34
SOIL (A)	0.77	**0.84**	**1.87**	6	4	4	4	4	2	2	3	2%	W35
GEOLOGY, EK SLOPE (A)	0.73	**0.82**	**1.12**	5	4	3	3	3	2	3	3	2%	W36

Table 28 - Results for expert knowledge methodology – Needwood Forest
Note: For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Needwood Forest													
Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ > 0.2	C ξ > 0.4	C ξ > 0.5	C ξ > 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
NSI SLOPE (full weight)	0.00	**0.57**	**0.32**	11	1	1	1	1	-	-	1	43%	N19
EK SLOPE (full weight)	0.00	**0.57**	**0.32**	11	1	1	1	1	-	-	1	43%	N20
GEOLOGY	0.20	**0.46**	**0.63**	11	6	4	3	3	-	-	1	14%	N21
GEOLOGY, NSI SLOPE	0.21	**0.47**	**0.65**	11	6	5	3	3	-	-	1	14%	N22
GEOLOGY, NSI SLOPE, SOIL	0.39	**0.59**	**1.11**	11	8	6	5	3	-	-	1	7%	N23
GEOLOGY, SOIL	0.39	**0.59**	**1.11**	11	8	6	5	3	-	-	1	7%	N24
NSI SLOPE, SOIL	0.42	**0.60**	**1.18**	11	8	6	6	4	-	-	1	7%	N25
SOIL	0.42	**0.60**	**1.18**	11	7	6	6	4	-	-	1	7%	N26
GEOLOGY, EK SLOPE	0.21	**0.47**	**0.65**	11	7	5	3	3	-	-	1	14%	N27
NSI SLOPE (full weight) (A)	0.00	**0.93**	**0.38**	7	1	1	1	1	1	1	5	7%	N28
EK SLOPE (full weight) (A)	0.00	**0.93**	**0.38**	7	1	1	1	1	1	1	5	7%	N29
GEOLOGY (A)	0.53	**0.92**	**1.08**	5	4	3	3	3	1	3	5	0%	N30
GEOLOGY, NSI SLOPE (A)	0.53	**0.92**	**1.08**	6	4	3	3	3	1	2	5	0%	N31
GEOLOGY, NSI SLOPE, SOIL (A)	0.49	**0.68**	**1.22**	7	7	5	5	5	-	4	2	0%	N32
GEOLOGY, SOIL (A)	0.49	**0.68**	**1.21**	7	7	5	5	5	-	4	2	0%	N33
NSI SLOPE, SOIL (A)	0.53	**0.70**	**1.27**	7	7	5	5	5	-	4	2	0%	N34
SOIL (A)	0.52	**0.69**	**1.29**	9	6	5	5	5	-	2	2	3%	N35
GEOLOGY, EK SLOPE (A)	0.46	**0.91**	**1.10**	7	5	3	3	3	1	1	5	3%	N36

Table 29 - Results for expert knowledge methodology – Yeovil
Note:. For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Yeovil													
Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ > 0.2	C ξ > 0.4	C ξ > 0.5	C ξ > 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
NSI SLOPE (full weight)	0.07	**0.31**	**0.13**	16	4	3	1	-	-	-	1	35%	Y19
EK SLOPE (full weight)	0.06	**0.30**	**0.12**	16	3	3	1	-	-	-	1	34%	Y20
GEOLOGY	0.48	**0.58**	**1.54**	16	10	7	6	5	-	-	1	6%	Y21
GEOLOGY, NSI SLOPE	0.45	**0.55**	**1.40**	16	11	7	6	5	-	-	1	4%	Y22
GEOLOGY, NSI SLOPE, SOIL	0.46	**0.55**	**1.45**	16	11	8	6	4	-	-	1	8%	Y23
GEOLOGY, SOIL	0.45	**0.54**	**1.43**	16	11	8	6	4	-	-	1	8%	Y24
NSI SLOPE, SOIL	0.49	**0.59**	**1.19**	16	9	7	5	3	-	-	1	9%	Y25
SOIL	0.49	**0.59**	**1.20**	16	9	7	5	3	-	-	1	9%	Y26
GEOLOGY, EK SLOPE	0.48	**0.58**	**1.45**	16	11	7	5	5	-	-	1	8%	Y27
NSI SLOPE (full weight) (A)	0.13	**0.72**	**0.32**	12	3	2	1	1	1	1	5	11%	Y28
EK SLOPE (full weight) (A)	0.09	**0.66**	**0.22**	11	2	2	1	1	1	1	6	12%	Y29
GEOLOGY (A)	0.52	**0.61**	**1.69**	14	9	8	6	5	-	1	3	4%	Y30
GEOLOGY, NSI SLOPE (A)	0.55	**0.70**	**1.59**	13	9	7	5	5	1	3	2	2%	Y31
GEOLOGY, NSI SLOPE, SOIL (A)	0.56	**0.65**	**1.71**	12	8	7	6	5	-	2	4	4%	Y32
GEOLOGY, SOIL (A)	0.54	**0.63**	**1.67**	12	7	7	7	6	-	3	3	6%	Y33
NSI SLOPE, SOIL (A)	0.53	**0.63**	**1.23**	13	7	7	6	3	-	2	3	6%	Y34
SOIL (A)	0.53	**0.62**	**1.25**	14	7	7	6	3	-	2	2	8%	Y35
GEOLOGY, EK SLOPE (A)	0.51	**0.62**	**1.46**	14	9	6	6	5	-	2	2	5%	Y36

6.10 Expert knowledge methodology discussions

The expert knowledge methodology innovatively built pseudo-quantitative model inputs based upon qualitative data contained primarily within published literature. The use of expert knowledge of the relationships between geology and parent material led to consistent increases in map value (ψ_3) across all areas for the unamalgamated tests over those of the data dictionary methodology (compare red and blue bars in Figure 37).

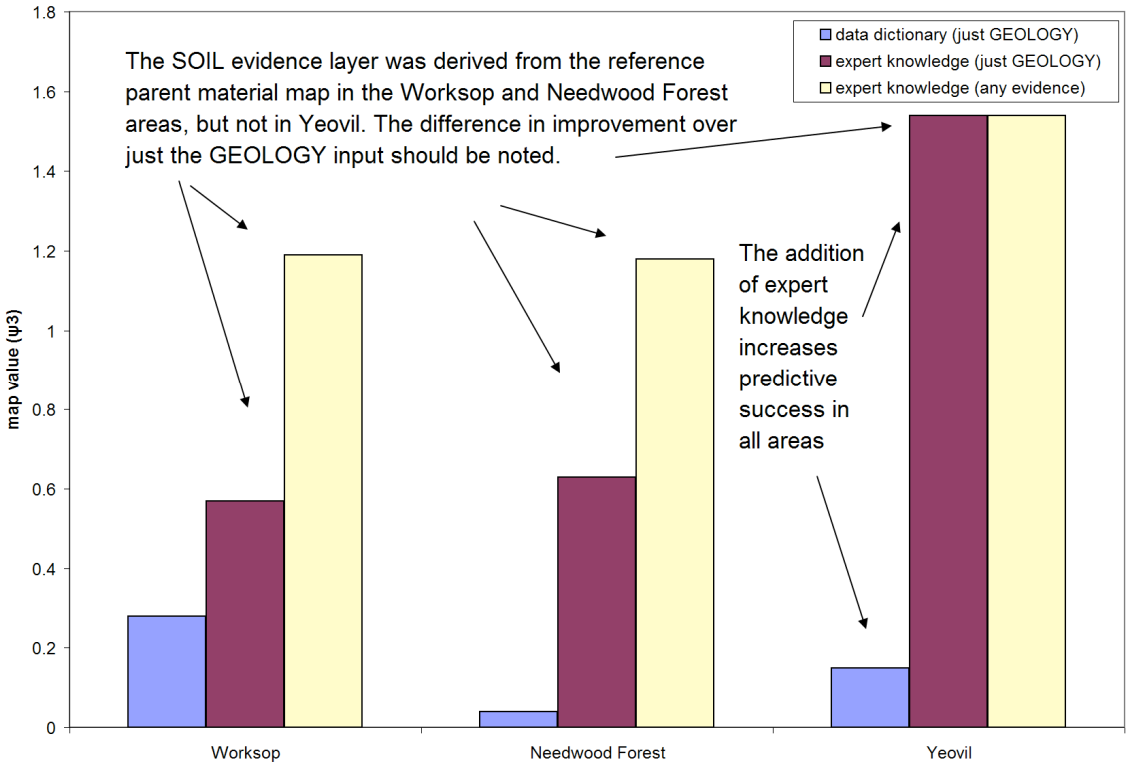


Figure 37 – Comparison of the data dictionary and expert knowledge methodologies (tests with no amalgamated units)
Note: the data dictionary tests use the surface geology input and the NSRI classification

This improvement was brought about by two items of knowledge. Firstly, a knowledge of the parent materials in the local area obtained from the Soil Records enabled the restriction of the number of predicted parent material units. Therefore, this dropped from 96 to a maximum of 16. Secondly, the improved lithological and historic terminology information obtained from the detailed BGS Lexicon, and the knowledge of relationships between soil types and geological units obtained from the Soil Records

enabled more accurate definition of the relationship between soil parent material and the surface geology.

When additional evidence layers were added, further increases in map value were achieved in the Worksop and Needwood Forest areas (compare yellow and red bars in Figure 37). In the Yeovil area, the GEOLOGY only input (Test Y21) achieved the highest unamalgamated map value.

Compared to the data dictionary methodology, in the expert knowledge method, when class amalgamation was used to correct misclassification, smaller increases in map value were achieved (Figure 38). In the Needwood Forest area, lower values were achieved. This anomaly is discussed in more detail later.

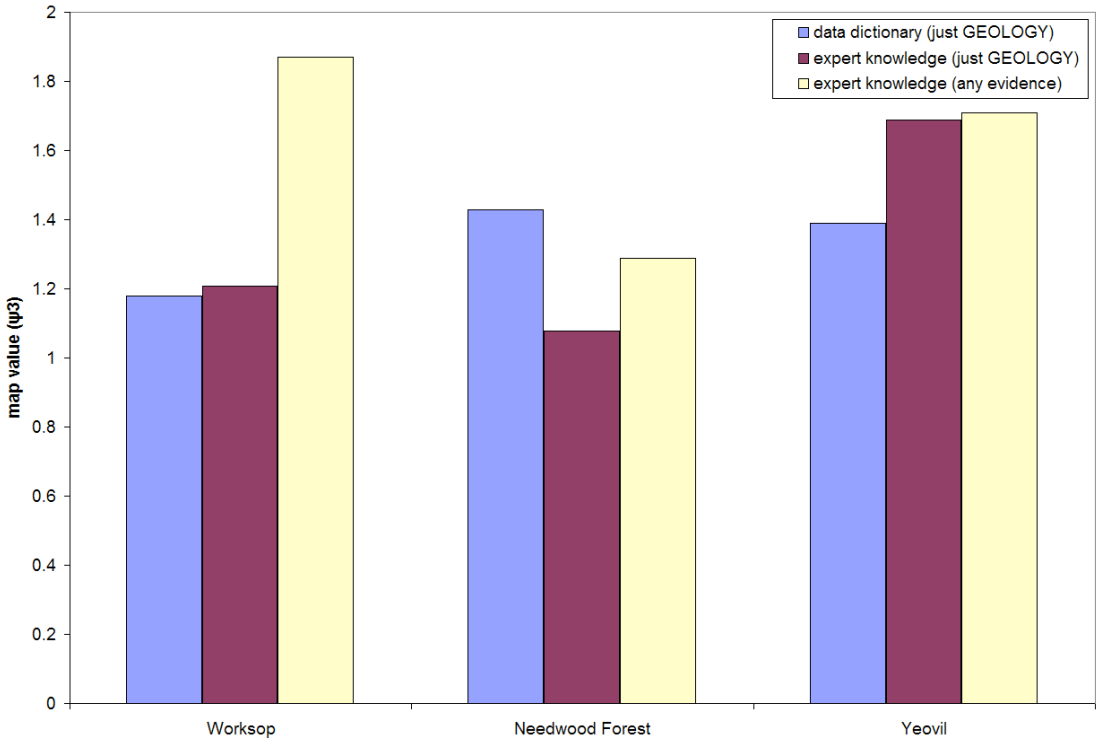


Figure 38 - Comparing the success of the data dictionary and expert knowledge methodologies (tests with amalgamated units)
Note: the data dictionary tests use the surface geology input and the NSRI classification

In all three study areas, the addition of the SOIL layer allowed higher map vales to be achieved than using just the GEOLOGY input in this method. It is clear that the SOIL input, which was created from the detailed reference map in the Worksop and

Needwood Forest areas, has a strong influence on the positive prediction of parent material in these areas. While a contributing factor in the Yeovil area, the SOIL evidence layer is not so dominant.

Table 30 – Comparison of the results with the highest map values (ψ_3) from the first two methodologies.
Note: Map value (ψ_3) is presented in bold, the test number in brackets and the inputs are listed. Note, in the data dictionary methodology, the only input used was GEOLOGY. The lower map value achieved in Needwood Forest using the expert knowledge method results from a greater flexibility of the first method, when combined with amalgamation. Here, numerous incorrectly identified classes were amalgamated with the correct classes (previously unpredicted in the data dictionary initial translation) leading to a higher map value.

Study Area	Data Dictionary	Expert Knowledge
Worksop	**1.25** (W6) GEOLOGY	**1.87** (W35) SOIL
Needwood Forest	**1.43** (N3) GEOLOGY	**1.29** (N35) SOIL
Yeovil	**1.39** (Y3) GEOLOGY	**1.71** (Y32) GEOLOGY, SOIL, SLOPE

The success of the evidence layers (GEOLOGY, SLOPE, SOIL) at predicting the various parent material classes are examined in detail in Appendix 6. It is evident that some evidence layers are more successful at predicting certain parent material classes than others. Additionally, a strong correlation is seen between the extent of the parent material class in the area and the class values (ξ) achieved. The relationship between extent, class value and the success of the predictors and parent material classes is discussed in more detail in section 8.6.3.

The relative success of the individual evidence layers will now be discussed after which comments will be made on the combination of these layers.

6.10.1 Expert Knowledge SLOPE and NSI SLOPE Inputs

Of the possible landscape attributes which could predict parent material, such as elevation, aspect, curvature etc., the only one with consistent available expert knowledge was slope. Even so, the results from this methodology indicate that expert knowledge of slope proved to be a very poor predictor of soil parent material in all three study areas (see Tests W20, N20 and Y20) compared with the other evidence layers. Figure 39 compares the highest map values (ψ_3) from this methodology (red and pink bars) with those obtained from models using slope inputs derived from expert knowledge (green bars) and the national NSI dataset (blue bars) for both amalgamated and unamalgamated tests.

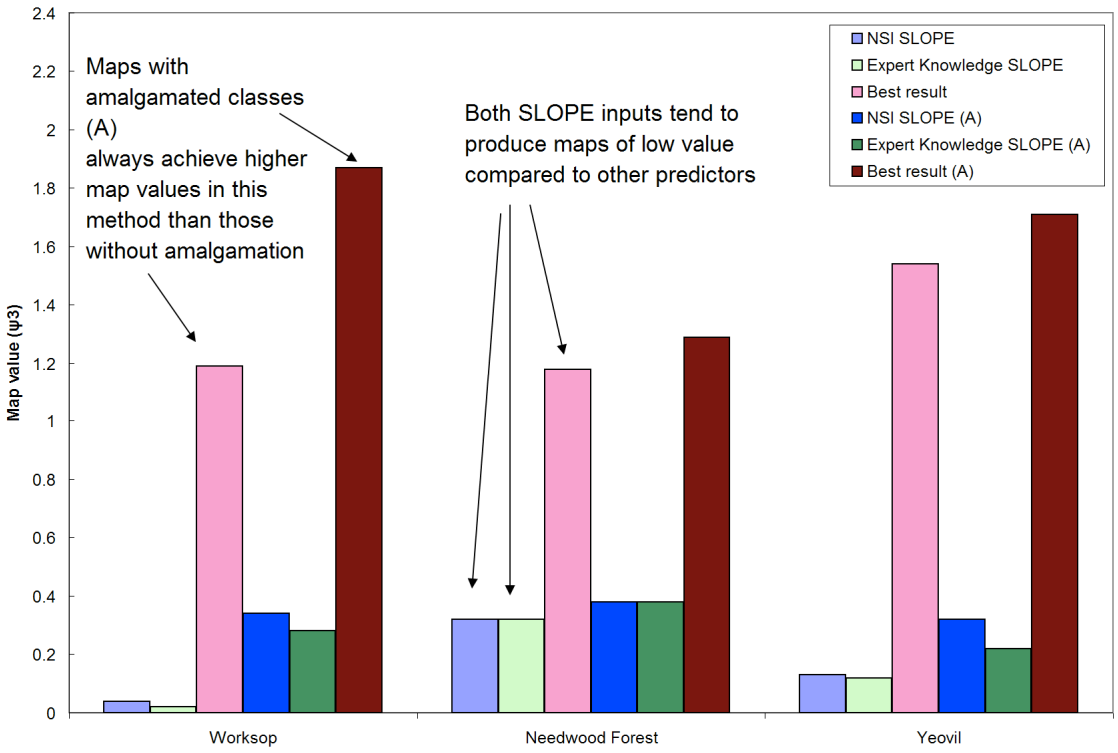


Figure 39 - Comparison of the map values (ψ_3) achieved by NSI and Expert Knowledge (EK) SLOPE inputs
Note: (A) indicates maps with amalgamated classes. The 'Best result' is that which is achieved using any combination of inputs in this methodology for that study area.

The alternative NSI slope dataset was derived from 5,148 NSI points recording parent material and slope, across all of England and Wales. While the NSI slope input tends to perform marginally better than the expert knowledge slope input, both are very poor,

and, when combined with other datasets tend to achieve maps of lower value than those achieved without the slope input.

As shown in Table 31, in the Worksop area, one of the failings of the NSI slope evidence layer as a predictor is the over-prediction of limestone (BhB1 and BhB2). These units are predicted as most likely by the model for all slope classes except steep slopes, where sandstones are predicted.

Comparing the probability and most likely parent materials between the two slope models revealed that the evidence layer based on expert knowledge resulted in a more diverse map than the NSI slope input. The expert knowledge slope predicted five of the nine parent material types present in the Worksop area, compared with the three from the NSI input.

Additionally, the probability values for the expert knowledge slope input were higher (Table 31). With the expert knowledge slope input, probabilities (P(H|E')) are commonly above 0.4, while for NSI they rarely achieve a higher value than 0.3. This more certain input arises from the use of the qualitative statements from the Soil Records which were relatively few compared to the sample of 5,148 points for NSI. Being a national dataset, the NSI input provides a more even distribution of parent material classes across a wider range of slopes.

Table 31 – The parent material classes which are predicted by expert knowledge derived from soil records and the NSI Slope evidence layer for the Worksop area.
Note how the NSI slope is less diverse predictor than expert knowledge

	Expert Knowledge Slope		NSI Slope	
Slope Class	**P(H|E')**	**Most likely P.M.**	**P(H|E')**	**Most likely P.M.**
level	0.508	BhB1 (limestone)	0.254	BhB1 (limestone)
gentle	0.293	EiE1 (drift)	0.247	BhB1 (limestone)
moderate	0.451	FiF1 (clay / mud)	0.285	BhB1 (limestone)
strong	0.494	BoB2 (sandstone)	0.280	BhB1 (limestone)
moderately steep	0.761	BoB2 (sandstone)	0.295	BhB1 (limestone)
steep	0.911	BhB2 (limestone)	0.277	BoB2 (sandstone)
very steep	1.000	BhB2 (limestone)	0.312	BhB2 (limestone)
precipitous	0.634	BhB2 (limestone)	0.285	BhB2 (limestone)

As the limestones (BhB1 and BhB2) are the most common parent material types in the Worksop area, the over prediction by the NSI model helps achieve a slightly higher map value than that achieved by the expert knowledge input, derived from the published literature. There are a few exceptions, where slope enables a higher map value to be achieved, for example test N22 (GEOLOGY and NSI SLOPE) outperforms N21 (GEOLOGY only) due to the under prediction of "drift with siliceous stones" (EiE1) by the geological input and over prediction of this parent material by the slope datasets. Indeed, EiE1 was the only unit predicted by the slope model. Thus, while slope can help to produce maps with the highest map value, the improvements tend to be rare and very marginal (ψ_3 difference of 0.02 between N22 and N21). Furthermore, they tend to achieve this success by the over prediction of the most extensive unit in the study area (for comparison see N21 versus N22 (Table 28) and Y32 versusY32, (Table 29)).

6.10.1.1 GEOLOGY Surface Geological Map

The surface geological map, GEOLOGY, combined with expert knowledge, proved to be a far better predictor of parent material than either slope input. The expert knowledge obtained led to great increases in map value for unamalgamated tests (Table 32).

Table 32 - Comparison of map values (ψ_3) using the GEOLOGY surface geology dataset in the first two methodologies

Study Area	Data Dictionary (ψ_3)	Expert Knowledge (ψ_3)
Worksop	0.28 (W1)	0.57 (W21)
Needwood Forest	0.04 (N1)	0.84 (N21)
Yeovil	0.15 (Y1)	1.54 (Y21)

The most dramatic improvement brought about by the use of expert knowledge is in the Yeovil area (Y1 versus Y21) where there is an increase in ψ_3 from 0.15 to 1.54 (Figure 37). A visual comparison of these two models (Figure 40 and Figure 41) clearly demonstrates the improvement in the initial prediction of parent material brought about by the use of expert knowledge.

A key benefit of expert knowledge is the restriction of parent material classes to those which are known to be present in the area. Table 33 displays the translation from the surface geology unit to the most likely parent material for the data dictionary and expert knowledge methodologies for the Worksop area. Where there are differences in the classification, these are emboldened. While the predicted lithology of the parent material tends to be similar between the methodologies, the physical nature of the parent material is often unknown without expert knowledge.

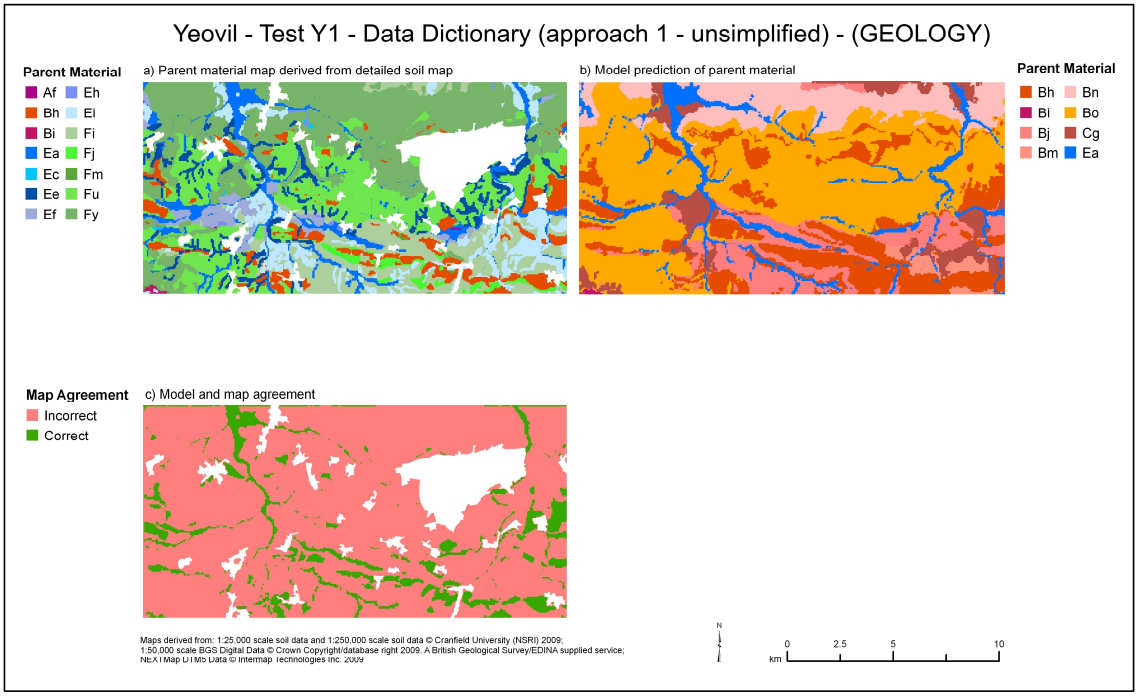


Figure 40 - Test Y1 maps (Data dictionary methodology, Approach 1)
Input: GEOLOGY (surface); Classification: NSRI PM_LITH; $\Psi_3 = 0.15$; $\theta_1 = 0.13$ $C_e = 3$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

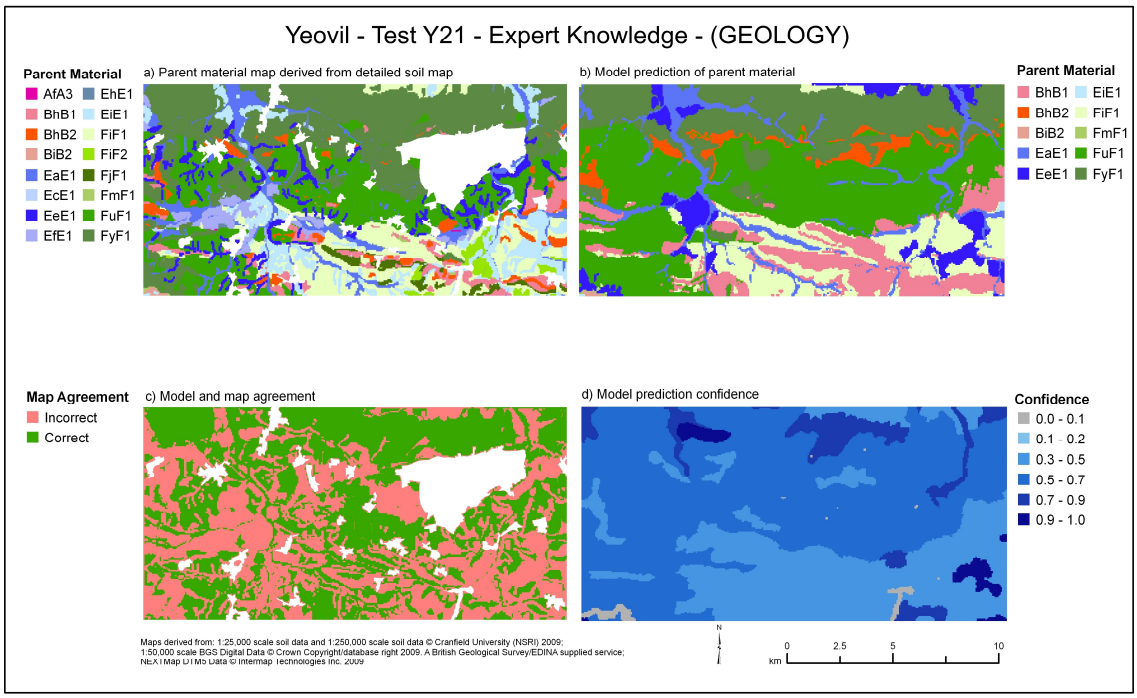


Figure 41 - Test Y21 maps (Expert knowledge methodology)
Input: GEOLOGY (surface); Classification: NSRI PARLITH; $\Psi_3 = 1.54$; $\theta_1 = 0.58$ $C_e = 10$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

Table 33 – Translation table resulting from the expert knowledge, compared with the parent materials assigned in the data dictionary methodology for Worksop (Test W21).
Note: P(H|E') is the model derived probability that the most likely parent material (PM) is correct. Where predictions are different between the methods, they are emboldened.

GEOLOGY UNIT	BGS Lexicon Description	Data Dictionary PM (1)	Expert Knowledge PM (2)	P(H|E')
ALV-CSSG	ALLUVIUM - CLAY, SILT, SAND AND GRAVEL	Ea (alluvium)	EaE1 (alluvium)	0.951
BTH-DOLM	LIMESTONE, DOLOMITIC	Bh (limestone)	BhB1 (limestone)	0.500
CDF-CAMD	CALCAREOUS MUDSTONE	Bj (mudstone)	**FiF1 (clay / mudstone)**	0.954
CDF-DOLO	DOLOMITE ROCK	Bh (limestone)	BhB1 (limestone)	0.500
EDT-CAMD	CALCAREOUS MUDSTONE	Bj (mudstone)	**FiF1 (clay / mudstone)**	0.593
EDT-MDSD	MUDSTONE AND SANDSTONE	Bm (clay / mud)	**FiF1 (clay / mudstone)**	0.376
EDT-SDST	SANDSTONE	Bo (sandstone)	BoB2 (sandstone)	0.958
GFDMP-SAGR	GLACIOFLUVIAL DEPOSITS - SAND AND GRAVEL	Cg (gravel)	**EiE1 (drift)**	0.965
HEAD-CSSG	HEAD - CLAY, SILT, SAND AND GRAVEL	Cg (gravel)	**EiE1 (drift)**	0.608
LNS-SDST	SANDSTONE	Bo (sandstone)	BoB2 (sandstone)	0.958
NTC-PEST	PEBBLY SANDSTONE	Bo (sandstone)	**EiE1 (drift)**	0.518
RTD1-SAGR	RIVER TERRACE DEPOSITS - SAND AND GRAVEL	Cg (gravel)	**EiE1 (drift)**	0.782
TILMP-DMSG	TILL, MIDDLE PLEISTOCENE - DIAMICTON, SAND AND GRAVEL	Cg (gravel)	**EiE1 (drift)**	0.965
TILMP-DMTN	TILL, MIDDLE PLEISTOCENE - DIAMICTON	Cg (gravel)	**EiE1 (drift)**	0.965
WBY-MDST	MUDSTONE	Bj (mudstone)	**BoB2 (sandstone)**	0.958

The unamalgamated tests consistently show improvement over the results obtained using the unamalgamated data dictionary methodology (Figure 37). Yet, in the amalgamated tests (Figure 38), using the GEOLOGY input, in the Needwood Forest area, a map of lower value (ψ_3 = 1.08, N30) was produced compared to its equivalent in the first, data dictionary methodology (ψ_3 = 1.43, N3, Figure 35). This is as a result of the flexibility of the first methodology to predict numerous parent material units which do not occur in the study area; for example, Bj and Cg (see Test N1 and N3, Table 6). With the later knowledge of the 'true' class brought about by the comparison with the reference map, these misclassified units can be accurately placed into the correct class. Even so, this example is the only one where the first methodology outperforms the expert knowledge methodology, and this is only achieved with later knowledge gleaned from the reference map. Furthermore, it must be mentioned that both these maps include very broad classes which cover almost the entire area (Figure 35, p126). This greatly reduces the usability of these maps. Therefore, extended discussion as to which of these maps is superior is unwarranted.

6.10.1.2 SOIL National Soil Map

The 1:250,000 scale National Soil Map (SOIL) was demonstrated to be a good predictor of soil parent material in all three areas, but especially the Worksop (W26, Figure 42) and Needwood Forest (N26) areas. This success was improved upon by guided amalgamation which increased overall accuracy (θ_1) by up to 22% at the expense of class detail (W30 (Figure 43) and N30).

While initially encouraging, on reflection, much of this success must be attributed to the relative age of the soil mapping. The Worksop and Needwood Forest areas were mapped in detail prior to the creation of the National Soil Map (SOIL) (Table 4, p44), and so these detailed sheets formed the basis of the simplified mapping used in the National Map. The Yeovil study area, however, was mapped in detail after the creation of the National Soil Map and so this area is more representative of what is likely to be expected in areas where no detailed parent material maps currently exist.

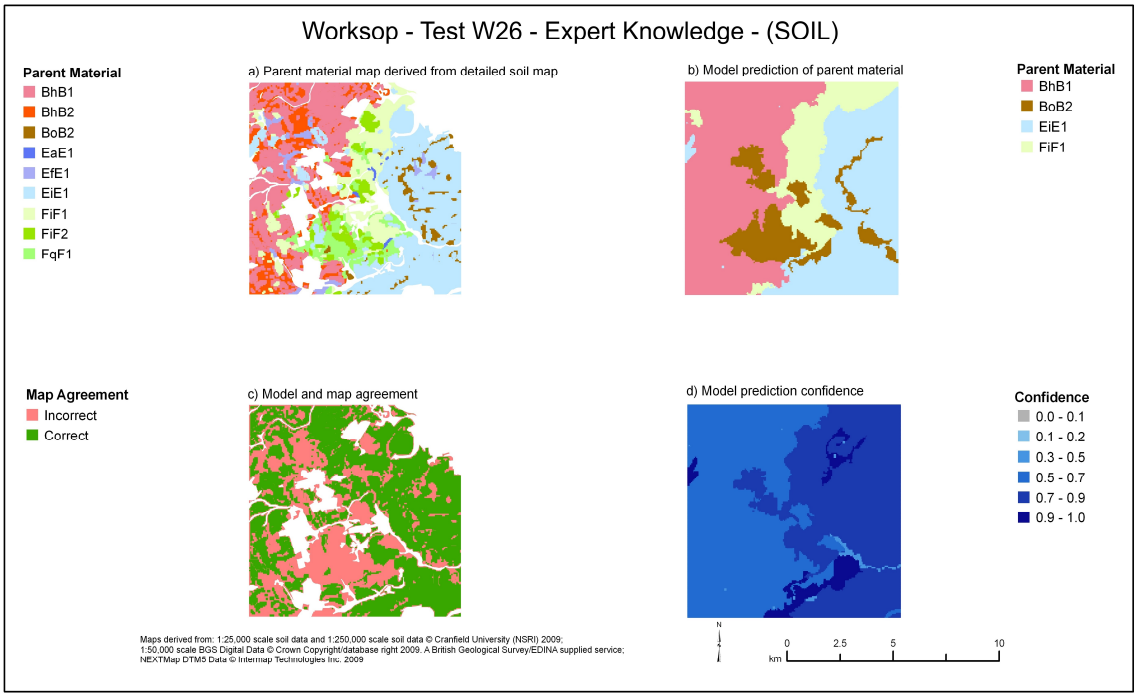


Figure 42 - Test W26 maps (Expert knowledge methodology)
Input: SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.19$; $\theta_1 = 0.62$ $C_e = 4$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

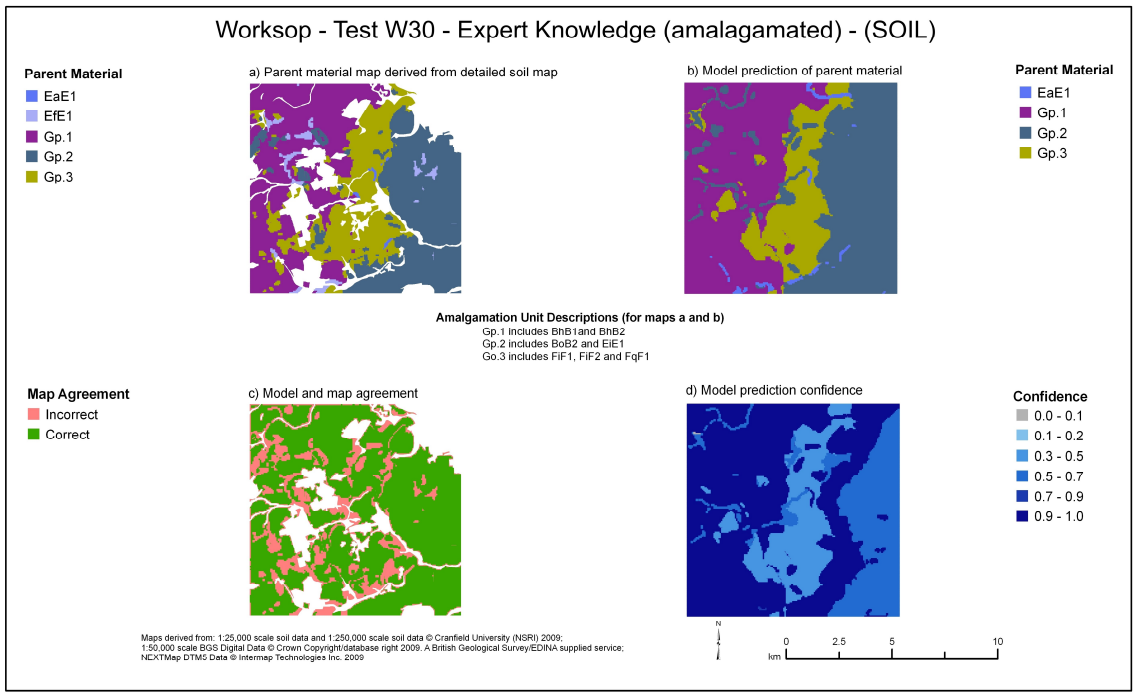


Figure 43 - Test W30 maps (Expert knowledge methodology)
Input: SOIL; Classification: Amalgamated NSRI PARLITH; $\Psi_3 = 1.87$; $\theta_1 = 0.84$ $C_e = 4$
A larger version is available in Appendix 4. NSRI are codes are described in Appendix 2.

While the SOIL input performed well in the Yeovil area (Test Y35, ψ_3 = 1.25), and certainly outperformed the SLOPE inputs (Tests Y28 &29, ψ_3 = 0.23 & 0.22), it did not match the map value obtained by the GEOLOGY input (Test Y30, ψ_3 = 1.69). A large proportion of this success is likely due to the extra detail in the linework of the 1:50,000 scale geological map compared to the 1:250,000 scale soil map, and the range of parent material units within the National Soil Map units.

6.10.1.3 Input layer combinations and the highest map values

It was hoped that the addition of multiple evidence layers would lead to higher map values. However, in the unamalgamated tests, the highest value maps were always generated with single inputs: SOIL for Worksop (W26) and Needwood Forest (N26) and GEOLOGY for Yeovil (Y21, Figure 41). It is an often accepted principal (Occam's razor), that theory should be no more complicated than required (Berger and Jefferys, 1991). This may offer some explanation why the added complexity of additional evidence layers did not necessarily increase the resulting map value in the unamalgamated tests. In the case of SLOPE, which performed particularly poorly, there appears to have have been the introduction of more noise than predictive evidence. Nevertheless, the use of multiple predicting layers tended to increase the number of effective classes. This arose due to the flexibility brought about by combining the probabilities of the different evidence layers. With a single evidence layer, the number of predicted most likely parent materials is limited by the number of classes in the evidence dataset (say, 9). With the addition of another evidence layer with, say 5, classes, there are now 45 potential class combinations and, theoretically, 45 possible most likely units.

With class amalgamation, a marginally higher map value (ψ_3 increase of 0.02) was achieved in the Yeovil area using all three inputs (Y32 ψ_3 = 1.71 (Figure 44)), over the amalgamated single GEOLOGY input (Y30, ψ_3 = 1.69) and at the cost of having 8 effective classes instead of 9.

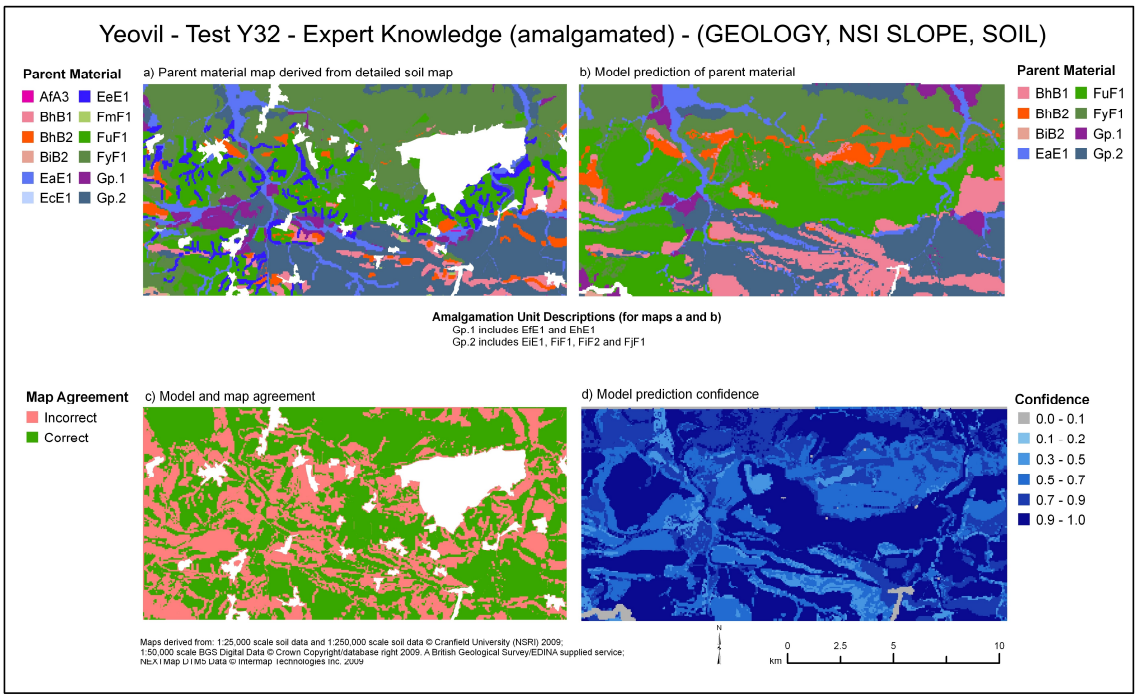


Figure 44 – Test Y32 maps (Expert knowledge methodology)
Inputs: GEOLOGY; NSI SLOPE, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.71$; $\theta_1 = 0.65$ $C_e = 8$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

6.10.2 Evaluation of the Expert Knowledge Methodology

Expert knowledge captured within published literature was identified for a range of environmental relationships. While many of these relationships were only partially described, consistent expert knowledge that was found to be of use included the relationships between the parent material and regional soil association and surface geology. Additionally, there was ample written expert knowledge on the relationship between slope and soil series (and hence, parent material). However, slope was shown to be a poor predictor of parent material when based on both expert knowledge and the national NSI dataset.

Compared to the data dictionary methodology, it was demonstrated that the use of expert knowledge generally led to more valuable maps with better defined parent material classes. This was particularly evident on tests with no class amalgamation.

Overall, the expert knowledge methodology achieved better results than the first, data dictionary methodology. There was considerable improvement in the initial unamalgamated translation from surface geology to soil parent material, as well as the additional information on the probability of each parent material class. This demonstrates that the expert knowledge captured in published literature represents a valuable, and previously unused, source of information which can greatly assist in the creation of rules to predict parent material. Future work needs to consider the scale and availability of this information, and the application of this knowledge beyond the study areas.

Because consistent misclassification remains with certain classes in the expert knowledge methodology, the process of guided amalgamation continues to increase map value. However, due to the restricted number of parent materials that were used in this methodology, there was less flexibility in the amalgamation to correct misclassifications. Thus, while more valuable maps were created in the Worksop and Yeovil areas using the amalgamated expert knowledge method, in Needwood Forest, because of this flexibility, a lower map value was achieved (cf. Tests N30; $\psi_3 = 1.08$ and N3; $\psi_3 = 1.43$).

The facility of combining multiple evidence layers did not greatly improve classification success over the use of just a single input. Nevertheless, this is an area which is worthy of further investigation, as it may be that the lack of quantitative data supporting the rules obtained from expert knowledge is limiting the success of the models.

While the defined relationships between parent material and the environmental covariates were improved, the lack of quantitative data on these relationships at the model building phase led to a great deal of uncertainty with regards to the probabilities which should be assigned to each evidence class – parent material class pairing. It is anticipated that, should quantitative data be available on the relationships between parent material and the environmental covariates (such as that obtained through data mining techniques), higher class values might be achieved.

Key points:

- Expert knowledge from published literature can be used to create inputs for probability models. These tend to produce more accurate predictions of parent material than one-to-one translations from geology maps.
- The use of multiple evidence layers enables the prediction of more parent material classes than can be achieved using just a geology map.
- GEOLOGY and SOIL tend to be better predictors of parent material than SLOPE these study areas
- The NSI SLOPE input based on a national quantitative sampling produced maps of equivalent value to the the EK SLOPE input based on local qualitative descriptions, although neither were particularly successful.

6.11 Recommendations

The following recommendation can be made from these findings:

- Future parent material mapping exercises should attempt to extract and incorporate into models expert knowledge held in published literature, with particular emphasis on the relationship between soil-types, their parent materials and geological units on which they form.
- Attempts should be made to better define the relationships between the geological units and the soil parent material units within the study area.
    - Specifically, can machine-learning characterise the relationships between soil parent material and geology, as well as other environmental layers which might be used as correlatives for soil parent materials? These correlatives might include:
        - Digital terrain model derivatives
        - Regional scale soil maps
        - Gamma radiometric remote sensing data
        - Electromagnetic remote sensing data
        - Thermal remote sensing data

7 DATA MINING METHODOLOGY

This chapter attempts to mirror that of the expert knowledge methodology, using the same probability model and evidence layers. However, instead of using qualitative expert knowledge to create the model inputs, an extensive, quantitative pairwise sampling on a 60 m grid of the relationships between parent material and the three environmental covariates (GEOLOGY, SLOPE and SOIL) is used to create the model inputs. Issues surrounding training and testing areas are briefly discussed, and the results of this quantitative method are compared with the previous methods.

7.1 Introduction

Data mining is an established technique, used extensively in many fields from the retail industry to environmental modelling. Standard data mining procedures consider a large number of datasets, and look at the patterns and relationships between the data members. However, this has not been done in this methodology where only pairwise relationships were considered. The data mining methodology was employed to discover to what extent parent material could be predicted using simple machine learning, but more importantly, how this fully quantitative technique compares to the other methods explored in this project. This methodology predicted parent material using the same probability model as the second, expert knowledge methodology, and using the same three evidence datasets. However the inputs to this model were not derived from the pseudo-quantification of published qualitative information but rather, by extensively sampling the relationships between the soil parent material and the same three environmental correlatives: surface geology map (GEOLOGY), National Soil Map (SOIL), and the slope model (SLOPE). Thus, this method provides an ideal comparison of the success of models trained on qualitative information and quantitative data.

7.2 The use of data mining in environmental models

Data mining involves the extraction of patterns or relationships from data. Data mining is a broad grouping of techniques, conceptually including supervised and unsupervised learning and classification (McBratney et al., 2003) using approaches such as regression trees and neural networks (Lark et al., 2007). Data mining offers a quantitative basis for assessing relationships between parent material and environmental covariates, although very few studies have actually attempted this. More common, however, is the application of such techniques to the prediction of other soil attributes (pH, carbon and clay content) and soil classes (e.g. Mayr et al., 2001; Moran and Bui, 2002; Bui et al., 2006).

Despite the prevalent use of data mining, many authors have expressed concerns surrounding the use of this type of machine learning. McBratney et al (2003) suggest that data mining of large databases of soil information should be approached with caution as the selection criteria of the soil sample sites is unknown, unless a systematic grid sample such as the NSI dataset is used. In a general paper discussing model formulation, Chatfield (1995) raises concerns regarding bias in data mining models, particularly conventional models when the model is assumed to be specified *a priori*, when actually it is based on data analysis. Lark et al. (2007) express reservations about the use of methods such as neural networks or decision trees unless a truly independent set of test data is available for validation. They, and other authors (Minasny et al., 2008) stress the ease and danger of over fitting data mining models to a training data set, as such models are not more widely applicable. While extrapolation of the models and techniques to other areas is beyond the scope of this research, these concerns do need some consideration.

In the last fifteen years, there has been a remarkable growth in volume and size of potential environmental evidence datasets, as organisations have digitised their existing maps, and GIS has become commonplace. It has been argued by Breiman (2001) that

because of the complexity of nature, a move should be made away from the traditional data modelling approach, where a physical mechanism which can be stated is assumed, and towards a black-box approach such as regression trees or neural networks where any mechanisms are unknown. Minasny et al., (2008) contend that scientific judgement should not be abandoned, yet concede that with the growth of datasets, previously unknown relationships can be discovered through black box and data mining approaches.

In this methodology, a simple form of pairwise data mining will be employed to quantitatively characterise the pairwise relationships between soil parent material and the three environmental covariates which have been previously used; GEOLOGY, SLOPE, and SOIL.

7.3 Assumptions

The following assumptions were made:

- That the 1:25,000 detailed soil maps accurately record the true distribution of soil type and soil parent material.
- That the distribution of soil parent material is related to the mapped surface geology, slope and National Soil Map.
- That the extent of the soil parent materials within the study area is known for the purposes of model generation.

7.4 Methods

The data mining methodology is very similar to that of the expert knowledge methodology (Figure 45). The main difference is the derivation of the evidence layer inputs to the probability model. In this methodology, they are based upon an extensive pairwise sampling of the relationships between parent material and the environmental covariates. Following the output of the parent material map, classes with consistent misclassification were amalgamated for a second suite of tests to improve map value.

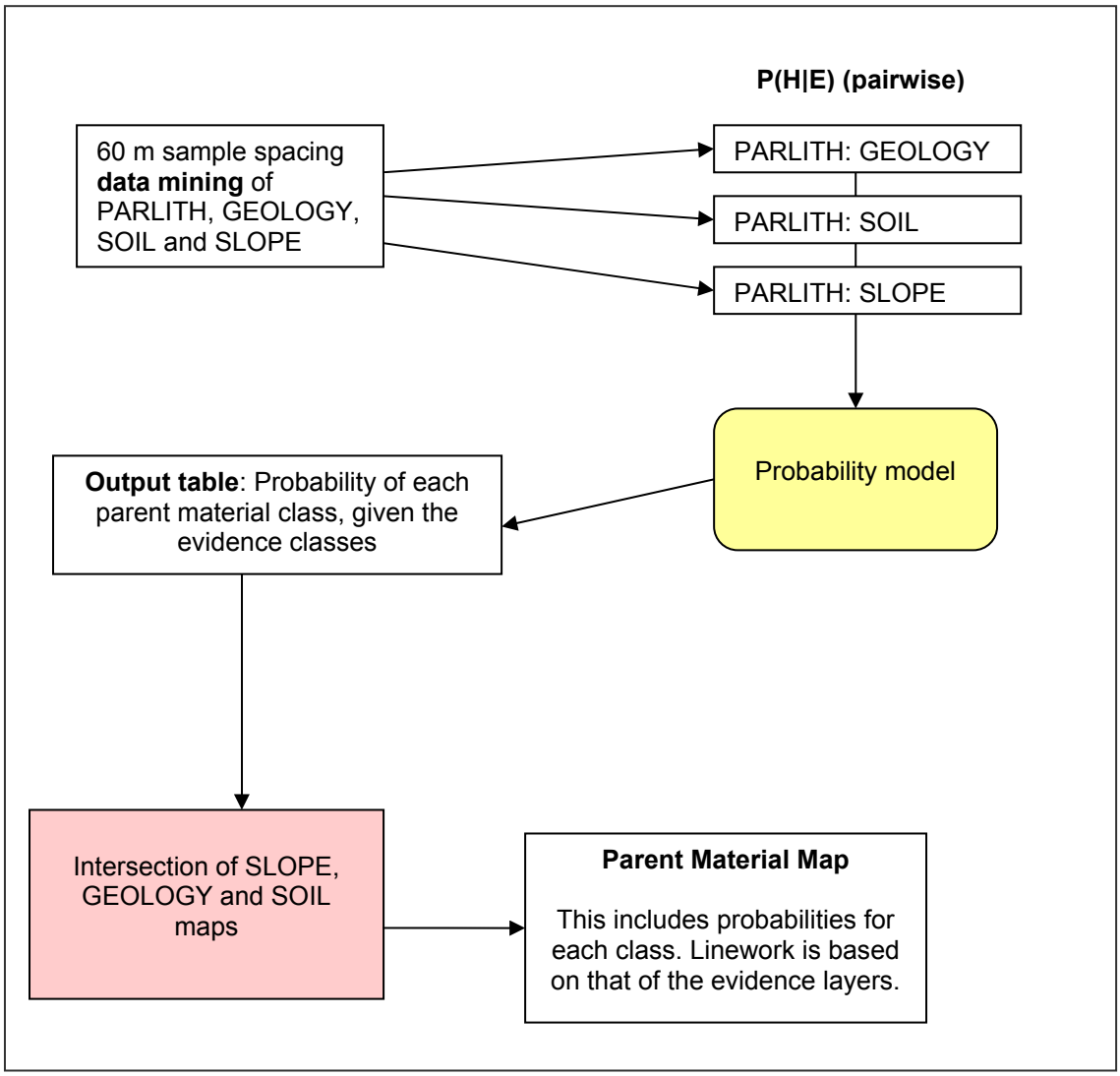


Figure 45 - The data mining methodology

7.4.1 Data sampling and evidence layers

Concerns were raised in the expert knowledge methodology that the probability model was based on qualitative data, and that this was difficult to quantify with any accuracy. It was felt that a fully quantitative approach, charactering the relationships between the environmental correlatives and the soil parent material would provide an interesting comparison. Therefore, a simple data mining exercise was used to populate the probability model with pairwise relationships between parent material and the environmental correlatives.

When training a model on data from a test area, it has been shown in research predicting soil that the more intensive the sample density, the more the model result will match the existing map (Moran and Bui, 2002). However, when data mining is used to predict an environmental variable on a previously unmapped area based on a training area, it is hypothesised that moderately spaced sampling will perform better than a very detailed sample or a very widely spaced sample. This is because a very detailed sample is likely to be over optimised for the local area, sampling local noise as well as the general geographic patterns. Such detailed patterns are less likely to apply beyond the training area. The Yeovil area was used to test this hypothesis. A standard test using the GEOLOGY and SOIL inputs were used.

Pairwise data mining was used to create probability model inputs. These models were trained using different sample densities on grids with 60, 100, 250, 370 and 590 m spacings. Irregular spacing intervals were chosen so selection of sample points would differ between the grid. It was found that if the models were trained and tested on the same area, a general increase in map value resulted from a smaller sample spacing (Figure 46). This trend is similar to those seen by Moran and Bui (2002) in relation to overall accuracy of predictions of a soil map increasing in line with increasing input data.

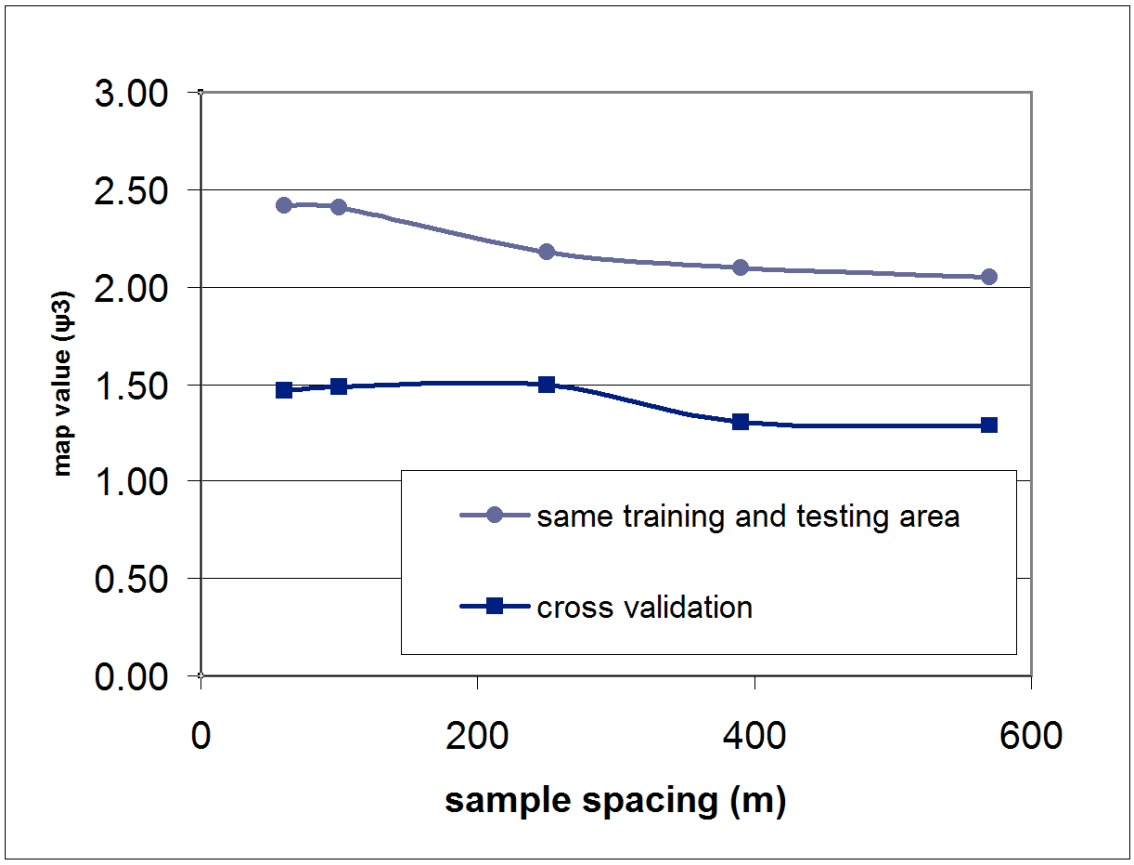


Figure 46 - Trend of map value with varying sample size
Note: cross-validation involved training one model on the western half, and testing it on the eastern half of the area, and vice versa. The two half-maps were then combined and analysed as normal. The model was tested on the Yeovil area with GEOLOGY and SOIL inputs.

When cross-validation was used to assess the performance of models trained and tested on adjacent areas, two key findings were made. Firstly, the map value achieved using the 60 m input decreased considerably compared to the model trained and tested on the same area (ψ_3 = 2.42 to 1.47; see Figure 46). This is the same level achieved by unamalgamated expert knowledge using the same inputs (Test Y24 ψ_3 = 1.43) and lower than the best unamalgamated test using expert knowledge (Y21 ψ_3 = 1.54).

Secondly, a very slight rise in map value was achieved over the cross validated 60 m sample by the 100 m and 250 m sample densities (ψ_3 = 1.47 to 1.49 and 1.50). This response was anticipated and probably results from the removal of noise, or spatial patterns less transferable to adjacent areas. As the sample spacing widen, the model has less data for training and so map value begins to drop. Nevertheless, with cross-

validation the change in map value is negligible between 60, 100 and 250 m grid sizes (Figure 46), and the change in overall accuracy (θ_1) and kappa (κ) is less than 0.1.

For the purpose of this research cross-validation has not be undertaken as the most challenging map value was sought to compare with the models derived from expert knowledge. Furthermore, as the expert knowledge was extracted from books and reports written about the whole area, testing with cross-validation was not necessarily comparable, and therefore the full 60 m was used to guide the quantitative model inputs used in this methodology.

This methodology used the same 60 m grid point shapefiles as used in the expert knowledge methodology. Each point was attributed with the surface geology (GEOLOGY), the slope class (SLOPE), and the soil association (SOIL), which is the broad soil class based on the National Soil Map. These samples were used for model building and analysis.

The model inputs for this methodology are discussed below, with general comments about these inputs being found in section 4.3.1 on p 65.

7.4.2 Joint probability tables – P(H,E)

The 60 m grid of attributed points was used to create probability distributions of the various evidence layers, for example, the probability of a point being a certain parent material (e.g. BhB1) given that the surface geology was limestone. These relationship tables were input as joint probability tables (e.g. P(H,E)) into the same probability models as used in the expert knowledge methodology.

7.4.3 Map purity tables – P(E,E')

Since no field work was carried out as part of this research, there was no empirical data for ascertaining the reliability of the class predictions in each evidence layer. For

example, when the map evidence (E) says that a point is BhB1, it is actually BhB1 in the field (E') 95% of the time. Without empirical data, subjective judgements were required. It was assumed that the GEOLOGY and SOIL classes would be correct 95% of the time, with the remaining 5% evenly distributed amongst the other classes. For the slope model, it was assumed that each slope class would be correct 80% of the time, with the final 20% split between the two most similar classes reflecting that most errors in the digital terrain model are unlikely to be extreme as this scale is continuous, not categorical. The precipitous slope class was given less confidence because of the poor quality of removal of some surface features. The choice of such values for map purity was pragmatic. It is known that maps and elevation are imperfect representations of reality, but without field survey like that undertaken to quantify the homogeneity of the soil units on the Harold Hill sheet (Sturdy, 1971) it was not possible to quantifiably populate the map purity tables.

After the main body of research was completed, a new technique of conveying the purity of the input layers was devised for situations such as those where there was limited information on the map purity. Instead of stating a likelihood of misclassification for each possible class pairing, a single value of confidence was supplied for each layer. This new assessment was not used in this study due to time restrictions, but is described in more detail in Farewell and Farewell (2010) and in Appendix 5. Initial analyses of this approach are encouraging and warrant further investigation.

7.4.4 Layer combinations and tests

The layers were weighted in an identical manner to the expert knowledge methodology. Each evidence layer was used as a single input with full weights (Tests 37, 38 and 43) and also combinations of the different evidence layers as described in Table 34. Where two or three evidence layers were used to predict parent material, for example, Tests 39-42, the probability model was used to combine the probabilities from each layer and

provide the most likely parent material based on the multiple evidence layers probabilities and weights.

Table 34 - Tests and weightings for the data mining methodology
Note: This table shows the weighting of evidence layers for the models for the unamalgamated tests The same weightings apply for the amalgamated tests 44 – 50. 1 indicates full weight. 0.5 indicates half-weighting.

		Test Number						
Evidence Layer	**Code**	**37**	**38**	**39**	**40**	**41**	**42**	**43**
NSI Slope	NSI SLOPE	1		0.5	0.5		0.5	
Surface Geology	GEOLOGY		1	1	1	1		
Soil Association	SOIL				1	1	1	1

7.4.5 Model outputs, data analysis and amalgamation

As in the expert knowledge methodology, the probability of each parent material was output for each point on the 60 m output grid. The most likely parent material was mapped and compared with the reference parent material maps derived from the detailed soil maps. Attempts were, once more, made to improve classification by class amalgamation using the guidelines described in section 5.4.4 (p 104). Identical analyses were performed on the model results for comparison with the first two methodologies. Confusion matrices were generated for each test, and the descriptive statistics, including class and map values (ξ, ψ_3) were calculated in each case. The most valuable maps were reported as the amalgamated tests in Tests 44 - 50 in each study area.

7.5 Results of data mining methodology

The results from the data mining methodology are presented in Table 35, Table 36 and Table 37.

Table 35 – Results for the data mining methodology – Worksop
Note: For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation;

Worksop Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
SLOPE (full weight)	0.05	**0.23**	**0.07**	9	**4**	4	-	-	-	-	**1**	28%	W37
GEOLOGY	0.47	**0.59**	**0.94**	9	**6**	3	3	3	**1**	-	**1**	12%	W38
GEOLOGY, SLOPE	0.42	**0.54**	**0.80**	9	**7**	4	3	3	-	-	**1**	7%	W39
GEOLOGY, SLOPE, SOIL	0.55	**0.65**	**1.58**	9	**8**	7	5	4	**1**	-	**1**	0%	W40
GEOLOGY, SOIL	0.55	**0.65**	**1.60**	9	**8**	7	5	4	**1**	-	**1**	0%	W41
SLOPE, SOIL	0.56	**0.66**	**1.61**	9	**8**	7	5	4	**1**	-	**1**	0%	W42
SOIL	0.54	**0.64**	**1.49**	9	**6**	6	4	4	**1**	-	**1**	22%	W43
SLOPE (full weight) (A)	0.11	**0.40**	**0.14**	5	**3**	3	3	-	-	3	**3**	2%	W44
GEOLOGY (A)	0.72	**0.81**	**1.32**	6	**4**	3	3	3	**2**	2	**3**	5%	W45
GEOLOGY, SLOPE (A)	0.76	**0.84**	**1.22**	5	**4**	3	3	3	**2**	3	**3**	0%	W46
GEOLOGY, SLOPE, SOIL (A)	0.76	**0.83**	**2.15**	7	**6**	6	5	5	**2**	2	**2**	0%	W47
GEOLOGY, SOIL (A)	0.76	**0.83**	**2.17**	7	**6**	6	5	5	**2**	2	**2**	0%	W48
SLOPE, SOIL (A)	0.78	**0.84**	**2.24**	7	**6**	6	5	5	**2**	2	**2**	0%	W49
SOIL (A)	0.78	**0.84**	**2.24**	7	**6**	6	5	5	**2**	2	**2**	0%	W50

Table 36 - Results for the data mining methodology - Needwood Forest
Note: For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation;

Needwood Forest Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
SLOPE (full weight)	0.00	**0.57**	**0.32**	11	1	1	1	1	-	-	1	43%	N37
GEOLOGY	0.19	**0.60**	**0.94**	11	6	5	3	3	-	-	1	14%	N38
GEOLOGY, SLOPE	0.57	**0.93**	**1.18**	11	4	4	3	3	1	2	5	0%	N39
GEOLOGY, SLOPE, SOIL	0.42	**0.62**	**1.45**	11	9	8	5	4	-	-	1	7%	N40
GEOLOGY, SOIL	0.42	**0.62**	**1.49**	11	8	8	5	4	-	-	1	7%	N41
SLOPE, SOIL	0.38	**0.63**	**1.14**	11	6	5	5	4	-	-	1	14%	N42
SOIL	0.40	**0.63**	**1.14**	11	5	5	5	4	-	-	1	14%	N43
SLOPE (full weight) (A)	0.00	**0.93**	**0.38**	7	1	1	1	1	1	1	5	7%	N44
GEOLOGY (A)	0.50	**0.91**	**1.21**	7	4	4	3	3	1	1	5	3%	N45
GEOLOGY, SLOPE (A)	0.50	**0.92**	**1.21**	7	4	4	3	3	1	1	5	3%	N46
GEOLOGY, SLOPE, SOIL (A)	0.51	**0.69**	**1.57**	9	8	7	5	5	-	2	2	3%	N47
GEOLOGY, SOIL (A)	0.51	**0.69**	**1.61**	9	7	7	5	5	-	2	2	3%	N48
SLOPE, SOIL (A)	0.56	**0.78**	**1.18**	8	5	4	4	4	1	2	3	3%	N49
SOIL (A)	0.54	**0.75**	**1.18**	9	4	4	4	4	1	2	2	7%	N50

Table 37 - Results for the data mining methodology – Yeovil
Note: For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation;
* indicates tests where class amalgamation could not increase map value (ψ_3).

Yeovil Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
SLOPE (full weight)	0.06	**0.26**	**0.09**	16	3	3	-	-	-	-	1	39%	Y37
GEOLOGY	0.55	**0.64**	**2.13**	16	9	9	7	6	-	-	1	9%	Y38
GEOLOGY, SLOPE	0.55	**0.64**	**2.10**	16	9	9	7	6	-	-	1	9%	Y39
GEOLOGY, SLOPE, SOIL	0.58	**0.66**	**2.29**	16	13	10	8	6	-	-	1	4%	Y40
GEOLOGY, SOIL	0.58	**0.66**	**2.42**	16	12	10	8	6	-	-	1	6%	Y41
SLOPE, SOIL	0.51	**0.61**	**1.37**	16	8	7	5	3	-	-	1	9%	Y42
SOIL	0.51	**0.62**	**1.37**	16	7	7	6	3	-	-	1	12%	Y43
SLOPE (full weight) (A)	0.05	**0.60**	**0.19**	12	2	2	1	1	-	1	5	15%	Y44
GEOLOGY (A)	0.58	**0.66**	**2.15**	15	9	9	7	6	-	1	2	6%	Y45
GEOLOGY, SLOPE (A)	0.58	**0.66**	**2.11**	15	9	9	7	6	-	1	2	6%	Y46
GEOLOGY, SLOPE, SOIL *	0.58	**0.66**	**2.29**	16	13	10	8	6	-	-	1	4%	Y47
GEOLOGY, SOIL *	0.58	**0.66**	**2.42**	16	12	10	8	6	-	-	1	6%	Y48
SLOPE, SOIL *	0.51	**0.61**	**1.37**	16	8	7	5	3	-	-	1	9%	Y49
SOIL *	0.51	**0.62**	**1.37**	16	7	7	6	3	-	-	1	12%	Y50

7.6 Discussions

The models driven by extensive pairwise data mining tended to produce more valuable maps than the previous two methodologies. Attempts were made to improve classification by class amalgamation. In this methodology, while the most valuable maps in each study area were created using class amalgamation, it did not commonly bring about the same level of improvement as seen in the previous two methodologies.

Without exception, the highest map values have been achieved using the data mining methodology (Table 38). This result comes with little surprise, as the entire test area was used to quantify the relationships used in this model. Consequently, it is likely these models were over-optimised for these areas. These results, therefore, represent the highest level of accuracy which is likely to be achieved using these datasets. As such, these results provide a useful comparison for those achieved with the other methods. The issue of the extent to which extrapolation is valuable is significant but demands considerable work, and as such, lies outside the scope of this current project.

Table 38 – The highest map values (ψ_3) from the first three methodologies.
Note: Map value (ψ_3) is presented in bold, the test number in brackets and the inputs are listed. In the data dictionary methodology, the only input used was GEOLOGY.

Study Area	Data Dictionary	Expert Knowledge	Data Mining
Worksop	**1.25** (W6) GEOLOGY	**1.87** (W35) SOIL	**2.24** (W50) SOIL
Needwood Forest	**1.43** (N3) GEOLOGY	**1.29** (N35) SOIL	**1.61** (N48) GEOLOGY, SOIL
Yeovil	**1.39** (Y3) GEOLOGY	**1.71** (Y32) GEOLOGY, SOIL, SLOPE	**2.07** (Y45) GEOLOGY

7.6.1 The most valuable maps

The most valuable maps using the data mining methodology were created using the National Soil Map (SOIL) input in the Worksop area ($\psi_3 = 2.24$, Test W50, Figure 47), the surface geology (GEOLOGY) in the Yeovil area ($\psi_3 = 2.07$, Y45) and both these layers in the Needwood Forest area ($\psi_3 = 1.61$, N48).

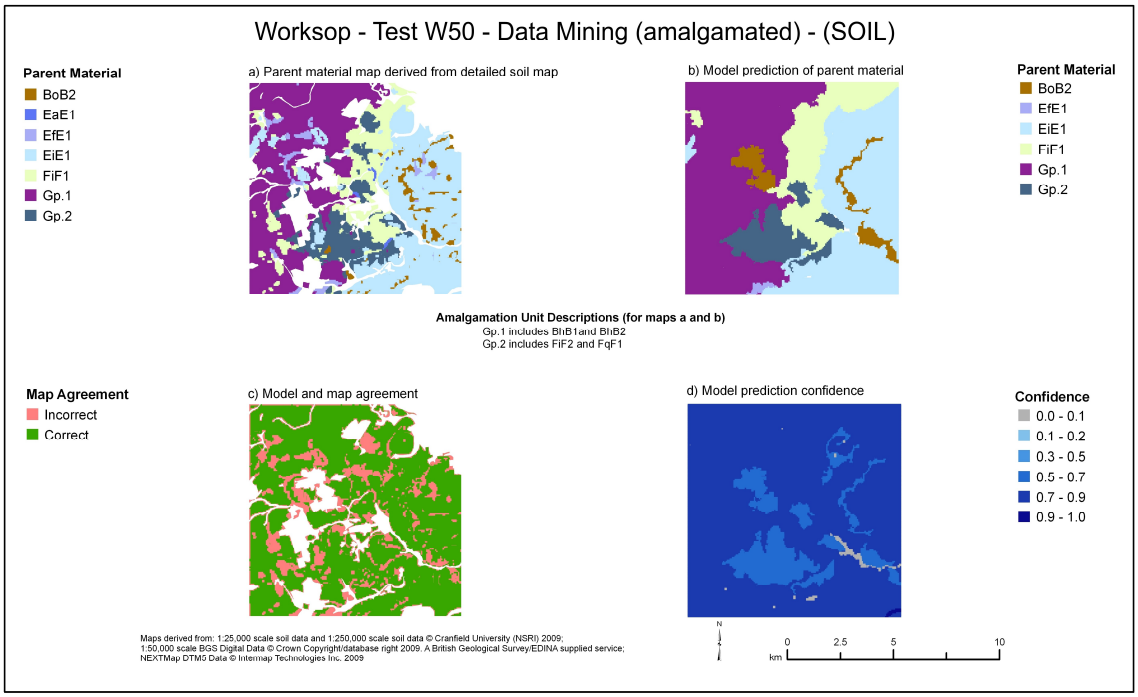


Figure 47 - Test W50 maps (Data mining methodology)
Inputs: SOIL; Classification: Amalgamated NSRI PARLITH; $\Psi_3 = 2.24$; $\theta_1 = 0.84$ $C_e = 6$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

SOIL remains a good predictor of soil parent material when using the data mining methodology. Increases in map value are achieved over the expert knowledge methodology in Worksop (ψ_3 increases from 1.19 (W26) to 1.49 (W43) in unamalgamated tests) and Yeovil (ψ_3 increases from 1.20 (Y26) to 1.38 (Y43)). However, in Needwood Forest, the national summary of soil series in each soil association (which are the map units in this layer), which was used as the expert knowledge for the SOIL layer, performs better than the data mined information on the local composition of the parent material within each map unit. In this case, ψ_3 marginally decreases from 1.18 (N26) achieved by the expert knowledge methodology

to 1.14 (N43) achieved by data mining. The reason for this is that the expert knowledge method predicts 7 classes, while the data mining method only predicts 6. When overall accuracy (θ_1) is considered, the data mining approach achieves an increase in the correct prediction of the most likely parent material from 0.60 (N26) to 0.63 (N43). So the data mining methodology achieves higher overall accuracy, but the expert knowledge method enables the prediction of more classes. A method of combining both advantages should be sought.

The National Soil Map (SOIL) still remains a better predictor of parent material than the surface geology (GEOLOGY) in both the Worksop and Needwood Forest areas. As discussed previously, this is likely to be due to the detailed soil map contributing to the National Soil Map in these areas. Therefore, it is likely that the linework of the National Soil Map more accurately depicts the local soil variation in these mapped regions than in areas unmapped at more detailed scales prior to the national mapping programme, such as the Yeovil study area.

When only the tests using guided amalgamation of surface geology driven models are considered (dashed lines in Figure 48), it is found that in the Needwood Forest area the data dictionary methodology outperforms both the expert knowledge and data mining methodologies

As discussed previously, this is due to the inaccurate initial translation of the parent materials in the Needwood Forest area, creating an abundance of classes with flexibility in the amalgamation. This aside, in every case, the unamalgamated tests (the solid lines on Figure 48) show a clear increase in map value over the previous methodologies as the data which is used to predict the parent material becomes increasingly detailed. Indeed, the most valuable map in the Yeovil area was created with this input (Y45, Figure 49)

.

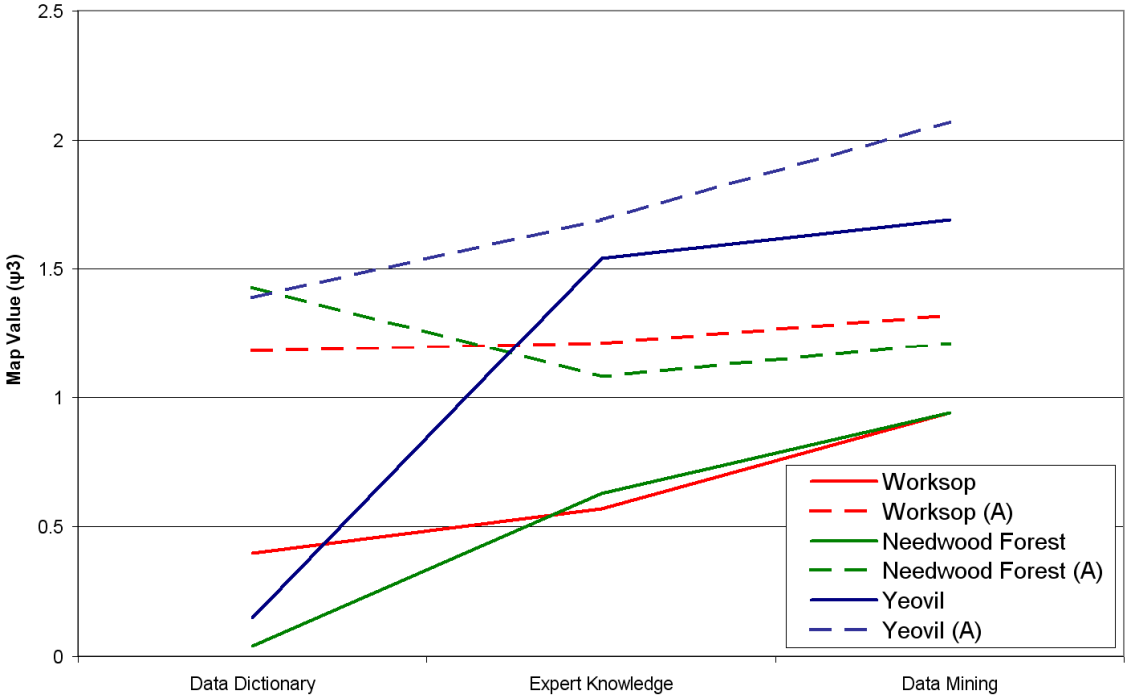


Figure 48 - Map value (ψ_3) comparison of the three methodologies using only the surface geology input (GEOLOGY).
Note: the dashed lines (A) represent the tests where class amalgamation was employed. Lines are used to display the trend and should not be interpreted as curves.

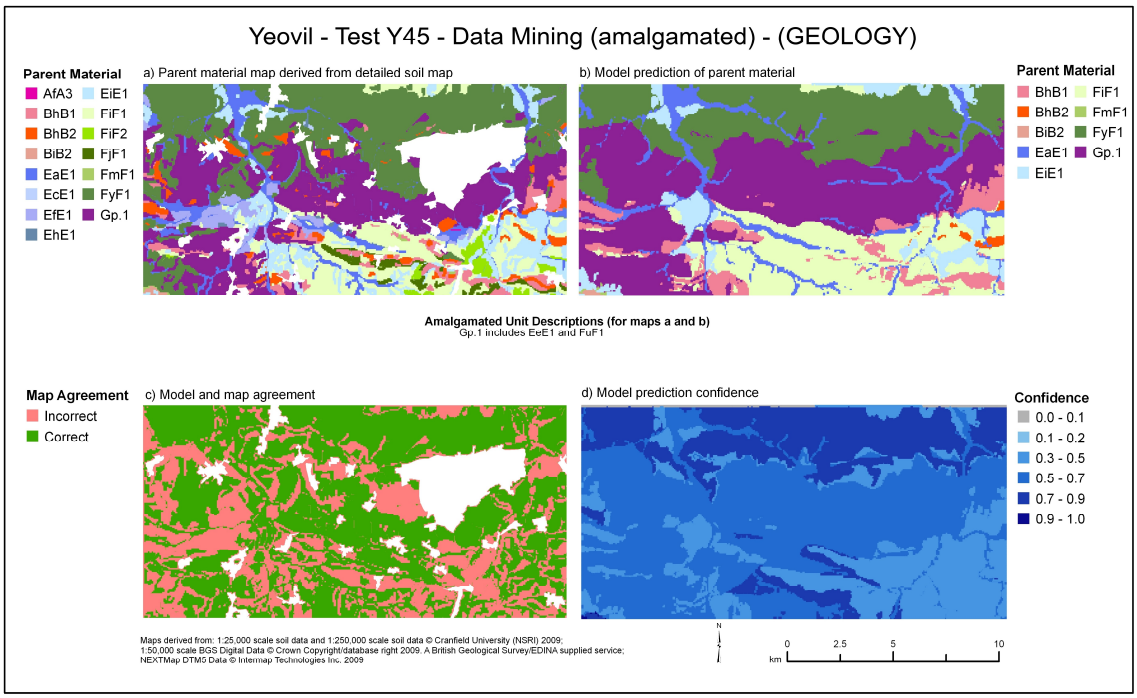


Figure 49 - Test Y45 maps (Data mining methodology)
Inputs: GEOLOGY; Classification: Amalgamated NSRI PARLITH; $\Psi_3 = 2.07$; $\theta_1 = 0.66$ $C_e = 9$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

SLOPE was also considered in this methodology. Even in the model trained on the whole study area, SLOPE was found to be a very poor predictor of soil parent material. Even with class amalgamation, low map values and few effective classes were achieved in all cases (see tests 37 and 44 for all areas, Table 35, Table 36 and Table 37). In the Yeovil area, the data-mined slope predictions of parent material were worse (Y44, ψ_3 = 0.19) than those predicted by expert knowledge derived from both qualitative descriptions in books (Y29, ψ_3 = 0.22) and the National Soil Inventory summary of slope and soil series (Y28, ψ_3 = 0.32). Nevertheless, none of these models produced very useful maps as they all contained a very broad class, with more than five amalgamated parent materials, which was very extensive, covering over 70% of the area.

7.6.1.1 Model success with multiple inputs

Except in the Needwood Forest area, where the most valuable map was created using both the GEOLOGY and SOIL inputs (Test N48, ψ_3 = 1.61, Figure 50), single evidence layer inputs achieved the highest map values when using the data mining methodology. It does not follow that the second and third highest values were achieved with the other two individual inputs, as these were generally achieved using multiple data inputs.

It is evident from Test N38 (Figure 51) that the surface geology (GEOLOGY) model over-predicts 'drift with siliceous stones' (EiE1) at the expense of thin drift deposits (FiF1 and FiF2), leading to a lower map value (ψ_3 = 1.21) than was achieved with both inputs (ψ_3 = 1.61). In particular, the combination of surface geology with the National Soil Map (SOIL) allows improved delineation of the 'clay or soft mudstone' units (FiF1, FiF2) which were poorly identified by the geological map. In situations such as these, multiple datasets prove valuable.

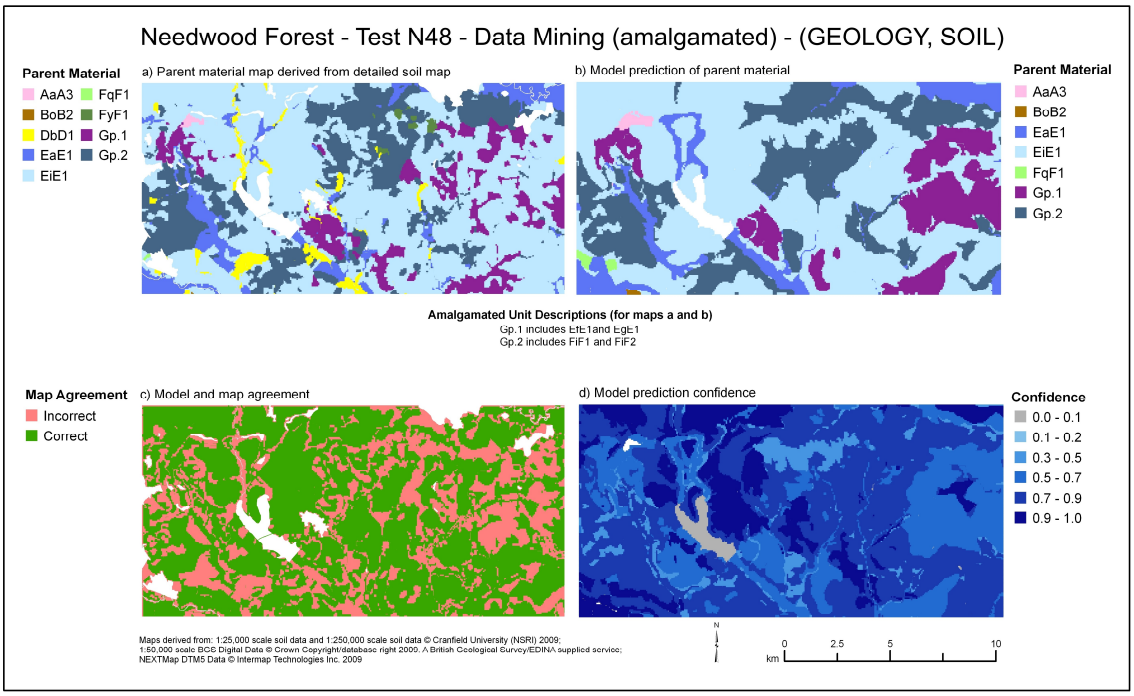


Fig . Resulting maps from Test N48 - Methodology: Data Mining; Inputs: DIGMAP, MUSID; Classification: Amalgamated NSRI; Summary Statistics: κ = 0.51 θ1= 0.69 ψ3 = 1.61
(κ = kappa coeffiecient, θ1 = overall accuracy, ψ3 = map value)

Figure 50 - Test N48 maps (Data mining methodology)
Inputs: GEOLOGY, SOIL; Classification: Amalgamated NSRI PARLITH; $\Psi_3 = 1.61$; $\theta_1 = 0.69$ $C_e = 7$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

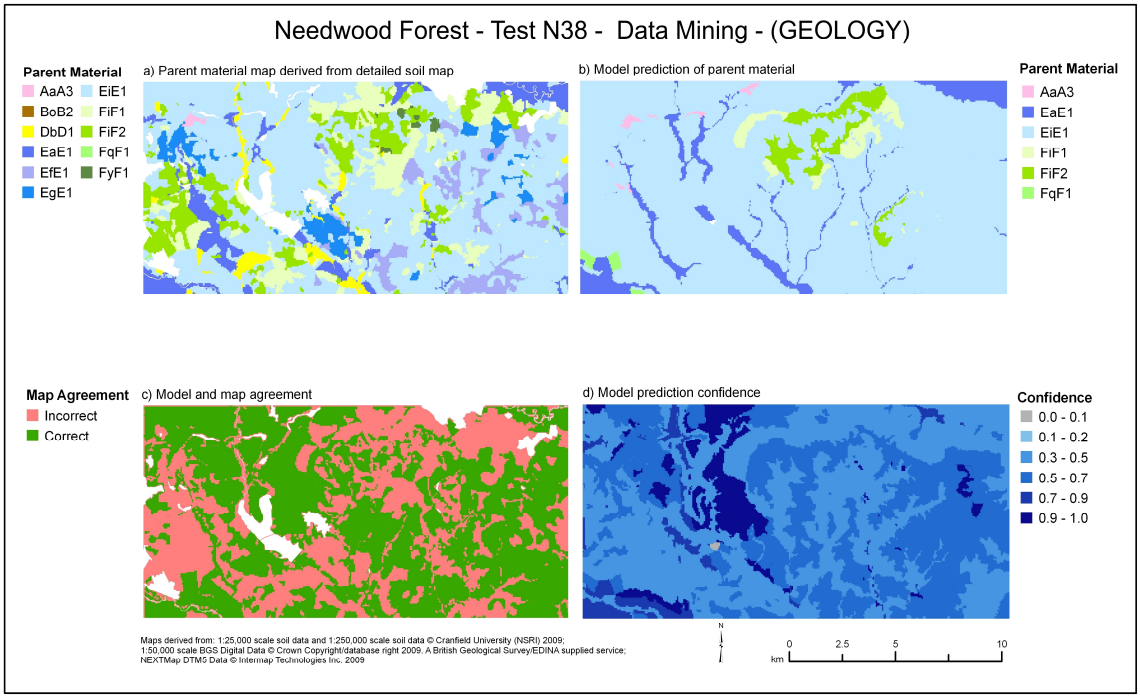


Figure 51 - Test N38 maps (Data mining methodology)
Inputs: GEOLOGY; Classification: NSRI PARLITH; $\Psi_3 = 0.94$; $\theta_1 = 0.60$ $C_e = 6$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

The quantity and class value of the different parent material units depends to a large extent upon the number of effective classes which are predicted by the models. All study areas show that as the number of evidence layers increase, the number of effective classes increase. This is the case in both the expert knowledge and data mining methodologies. It does not follow that these are the most valuable maps, but they do have the greatest class diversity. It is expected that if a multivariate sampling was undertaken, (SLOPE+GEOLOGY+SOIL vs. parent material), the number of effective classes would increase, however, the over fitting of the model to the training area would be even more noticeable.

7.6.2 Summary of the data mining methodology

This methodology produced the most valuable parent material maps to date, yet questions remain about the value of this approach beyond the sample areas. The close sample spacing (60 m) guiding the model inputs has resulted in high map values, yet such a density is unrealistic for application in a real world context. Furthermore, it is likely that the models resulting from this data are over optimised for the training areas, contain a certain level of 'noise' and are unlikely to perform as well in other areas. It has been demonstrated that models trained on one area and applied to an adjacent area do not perform as well as those trained and tested on the same area. The applicability of models derived from data mined inputs depends on how representative the training area is of the area to be tested. Further questions remain about the usefulness of slope as a predictor of soil parent material.

In real world situations, resources funding field work tend to be limited. Therefore, to reduce the amount of theoretical fieldwork to an achievable level, the effectiveness of wider, sparser, sample strategies should be investigated. Additionally, there is likely to be value in refining the model inputs derived from these sparser samples with rules gleaned from expert knowledge to create a combined expert system. Here the rules may be defined by expert knowledge, but quantified with assistance of sampling and data

mining. This would go some way to addressing concerns about the applicability of these models to adjacent areas.

Key points:

- The creation of model inputs using pairwise data mining on a dense sample grid created maps of high value, but the sample density is not pragmatic for application to new areas.
- Compared to GEOLOGY and SOIL, SLOPE appears to be a less effective predictor of parent material.

7.7 Recommendations

- The creation of model inputs from pairwise quantitative sampling produces parent material maps of high value and so quantitative data should be included where possible.
- Model inputs should be created from the data mining of much wider and more achievable sampling to consider the success of this approach with considerably fewer sample points.
- The qualitative information from expert knowledge should be combined with the quantitative data-mined information based on wider sample spacing, to create a harmonised rule set.

8 COMBINED EXPERT KNOWLEDGE AND DATA MINING METHODOLOGY

The data mining method used impractical, extensive sampling to predict parent material. This chapter investigates ways of creating parent material maps while reducing the need for extensive fieldwork. It describes the creation of model inputs using data collected on sparse sample grids of 700, 1400, 2100, and 2800 m. Firstly, tests are run using data from these samples in the same manner as the data mining methodology. Secondly, the model inputs from the expert knowledge methodology are combined with those from the sparse samples, creating new model inputs derived from both qualitative and quantitative information. Potential evidence layers to be created are first tested for association with parent material. Models using the chosen inputs are run and in discussion, the results are compared with previous methodologies, as are the values of the different parent material classes.

8.1 Introduction

In previous methodologies, it has been demonstrated that both qualitative knowledge and quantitative data can be used to produce model inputs for predicting soil parent material. Nevertheless, each of these approaches had limitations. The qualitative expert knowledge method lacked any quantitative framework, meaning assigned pairwise probabilities were uncertain. The quantitative data mining method used an unrealistically dense sample, and concerns remain about the applicability of the derived model beyond the training sample areas. This methodology investigates more pragmatic methods of creating parent material maps by combining aspects of each methodology, providing quantified expert knowledge inputs to the probability model.

As field survey is expensive, only the most essential can typically be undertaken. Therefore, a range of increasingly sparse sample strategies are investigated to ascertain appropriate sample densities for use in the landscapes under investigation. Soil parent material will be predicted using expert knowledge and four sparse sample grids (700, 1400, 2100 and 2800 m). The 700 m sample grid uses a similar number of points per km^2 to that undertaken for the mapping of the National Soil Map. As such, it is a more achievable level for reconnaissance survey than the 1 point every 60 m used in the data mining methodology. The wider spacings are multiples of the 700 m samplings, and are used to test level of detail required from field sampling to positively contribute to the parent material models.

8.2 The combined use of expert knowledge and data mining in environmental models

Many of the concerns raised about data mining and black box approaches (Lark et al., 2007; Chatfield, 1995; Minasny et al., 2008) as discussed in section 7.2, can be addressed with the used of a combined approach. For example, Minasny et al. (2008) express concern about the over-optimisation to the training area of models trained on data mining samples. In a probability model which uses pairwise relationships as the model inputs, the results of data mining can be examined on a pair-by-pair basis. Reference can be made to equivalent tables based entirely on expert knowledge, to ensure that such relationships are logical or explainable. This may improve the results achieved when such models are used to provide predictions beyond the training areas. It has also been noted that the extraction and formalisation of expert knowledge can be problematic because of poorly stated mental models (Lagacherie et al., 1995) which can be difficult to capture (McKenzie and Ryan, 1999). The use of quantitative data can highlight clear and consistent errors in the formalisation of expert knowledge.

(Minasny et al., 2008) suggest that hybrid methods between black box and data model approaches should be considered. They cite Bui et al. (2006) as a good example of a hybrid approach, where decision tree algorithms, derived from detailed examination of

data mined information were used to predict the environmental patterns. The algorithms were assessed and found to concur with known processes of soil formation.

An approach to integrating expert knowledge and quantitative data may be to individually run models based on both the qualitative knowledge and quantitative data based inputs. Once run, the resulting maps might be compared and individual units could be predicted by each model and then combined into a final map. However, there are some concerns with such an approach. Firstly, the resulting map may have gaps where no units are predicted, as perfect edge matching between the model runs is unlikely. Secondly, it is becoming apparent that the probability distribution within each class is likely to be of value to later environmental applications. Because the resulting maps contain a probability for each parent material class, it is unclear on what basis the selection of the best predicting model would be made. For example, would the selection be based entirely on the most likely class? In which case, other classes which may be almost as likely (e.g. see Table 9, p70) may not be as accurately predicted by the chosen methodology. For such reasons, this approach was not selected for use in this methodology.

Where empirical data was limited, Smith et al. (2007) supplemented the information with expert knowledge for input into a Bayesian belief network. A similar approach will be undertaken in this methodology, where expert knowledge will be combined and cross-verified with sparse data samples, prior to input into the probability models.

8.3 Assumptions

- That the 1:25,000 reference soil maps accurately record the true distribution of soil type and soil parent material.
- That the soil parent material is related to the mapped superficial and bedrock geology, and National Soil Map
- Expert knowledge (information from soil records and books) is available for an area adjacent to the study area
- The adjacent area for which expert knowledge is available is geologically similar, and have similar parent material / soil / geological relationships
- Expert knowledge is gathered a test area in this case, but it is assumed that this would be the same from the adjacent area, if available.
- That a systematic sample has been carried out recording soil parent material, geology, and the national soil map unit every 700 m, 1400 m, 2100 m and 2800 m across the study area.

8.4 Methods

This method has eight stages, which are outlined in Figure 52, and described in more detail below.

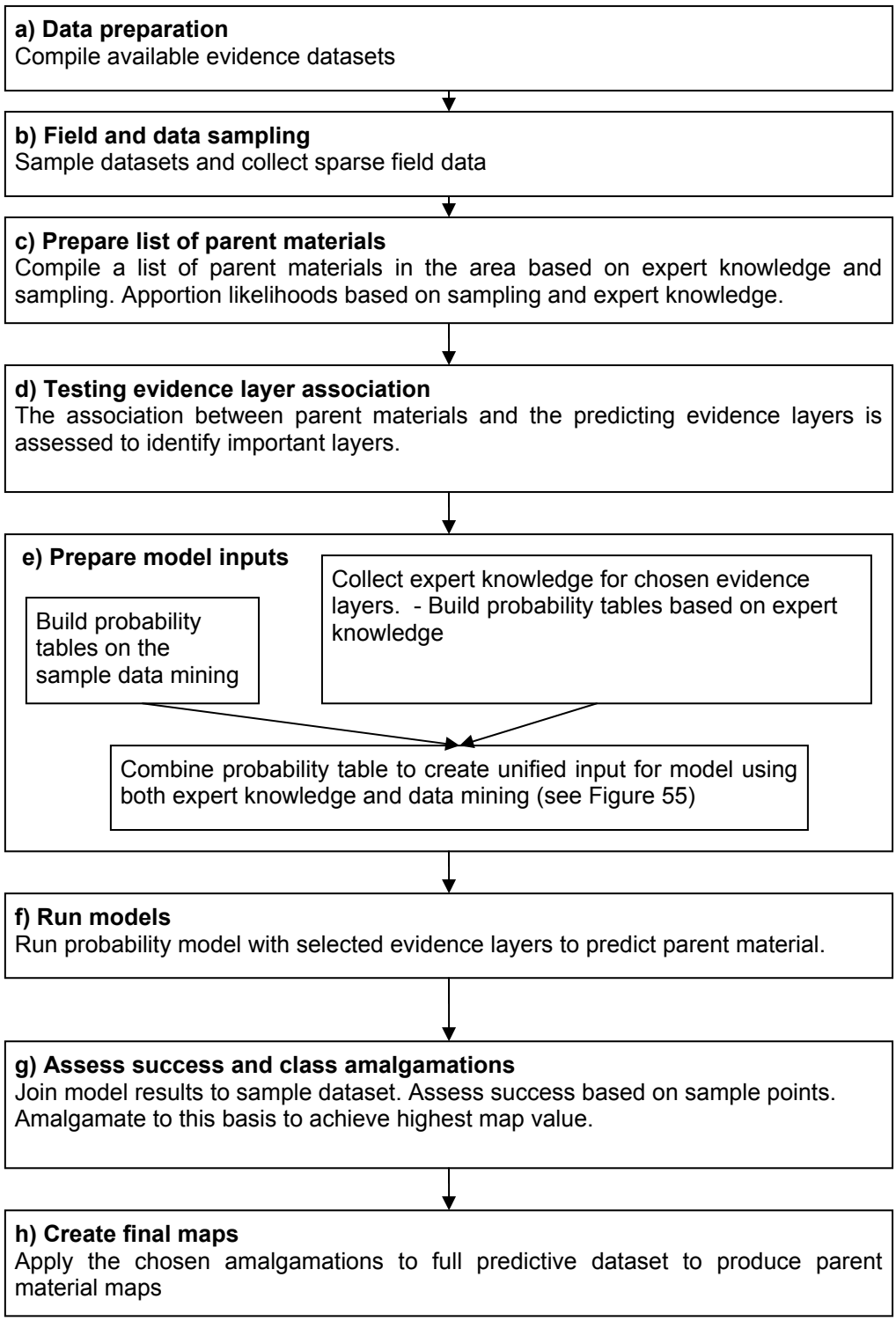


Figure 52 - Workflow for combined methodology

8.4.1 Data preparation (Figure 52a)

As in previous methodologies, three distinct geographic layers were selected as predictors of soil parent material: slope class map derived from a 5 m resolution DTM (SLOPE), 1:250,000 scale National Soil Map (SOIL) and a 1:50,000 geology map (GEOLOGY).

8.4.2 'Field' and data sampling (Figure 52b)

A systematic sampling of the data held of the study areas was carried out on four sample grids: 700 m, 1400 m, 2100 m and 2800 m. As field based sampling was not possible in this research, a sampling of the reference parent material maps was used. Each point is attributed with the class from the three evidence layers and also with the reference soil parent material class.

8.4.3 Preparation of the list of parent material classes (Figure 52c)

One of the purposes of the 'field' sampling was to determine the presence and abundance of parent materials that were likely to be present in the study area. Table 39 highlights the advantages of a grid survey for the task of characterising the parent materials which are likely to be present.

The true extent of parent materials in the Yeovil area (as obtained from the area of the polygons from the reference map) is compared to the extent of those parent materials estimated by three different methods:

1. the locations of auger bores in the LandIS database (augers, Figure 53a);
2. a 700 m systematic (grid) sample (Figure 53b);
3. the parent material extents estimated by the membership of soil series in the map units for the National Soil Map (SOIL, Figure 53c).

The sum of the differences in Table 39 is calculated according to Equation [6].

Table 39 – Comparison of the predicted extents of parent material units in the Yeovil area

	extent				difference		
parent material	true[1]	augers[2]	700 m grid[3]	SOIL[4]	augers	700 m grid	SOIL
AfA3	0.1%	0.3%			0.2%	0.1%	0.1%
BhB1	4.5%	7.8%	4.0%	6.1%	3.3%	0.5%	1.5%
BhB2	3.6%	3.8%	3.2%	4.2%	0.2%	0.5%	0.6%
BiB2	0.1%			0.0%	0.1%	0.1%	0.1%
BoB2 (N)				0.1%			0.1%
DaD1 (N)				1.9%			1.9%
EaE1	8.9%	5.3%	8.9%	2.4%	3.6%	0.0%	6.6%
EcE1	0.1%		0.3%		0.1%	0.1%	0.1%
EeE1	3.5%	2.3%	4.3%		1.2%	0.8%	3.5%
EfE1	2.4%	2.3%	2.6%		0.1%	0.2%	2.4%
EhE1	0.7%	0.8%	0.6%	0.2%	0.0%	0.2%	0.6%
EiE1	7.3%	2.8%	6.0%	1.8%	4.5%	1.2%	5.5%
FiF1	13.7%	13.6%	16.1%	14.3%	0.1%	2.4%	0.6%
FiF2	1.3%	2.0%	1.1%	4.1%	0.7%	0.2%	2.7%
FjF1	1.3%	0.8%	0.9%	2.5%	0.6%	0.4%	1.2%
FpF1 (N)				3.3%			3.3%
FqF1 (N)				28.7%			28.7%
FmF1	0.3%	0.8%	0.3%		0.5%	0.0%	0.3%
FuF1	23.0%	27.5%	23.6%	28.7%	4.4%	0.5%	5.7%
FxF1(N)				0.1%			0.1%
FyF1	29.0%	30.2%	28.2%	7.9%	1.2%	0.9%	21.2%
					sum of differences		
					20.7%	8.2%	86.8%

Note: 1) Extent based on parent material derived from the reference map. 2) Extent based on auger boreholes held in LandIS. 3) Extent based on 700 m grid sampling of data. 4) Extent based on soil association membership for units in the National Soil Map (SOIL). (N) represents units only predicted by SOIL. The difference is calculated by the absolute value of the prediction (e.g. augers, grid, or SOIL) minus the true value (see equation [6]).

$$\text{sum of difference} = \sum_{i=1}^{n} \left(| \, p - r \, | \right) \qquad [11]$$

Where p = predicted extent of parent material class; r = reference 'true' extent of parent material class for n parent material classes

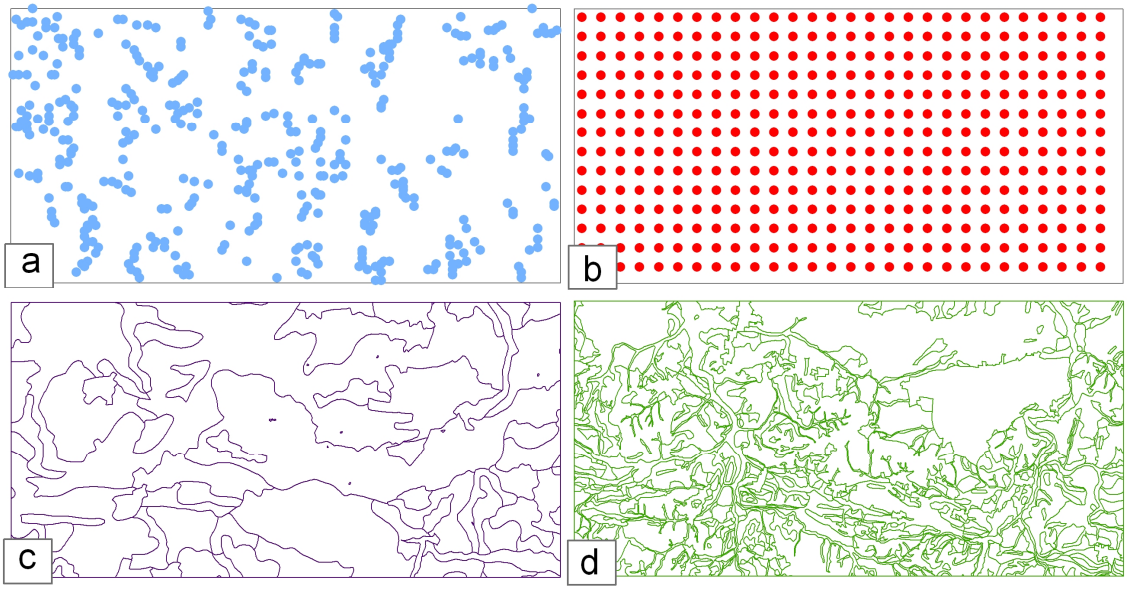


Figure 53 - A comparison of sample strategies in the Yeovil area
Note: compare with Table 39 - a) (augers) - Location of approximately 400 auger bores from LandIS database, collected for the mapping of the National Soil Map (SOIL). b) 700 m systematic grid (approximately 400 points) c) linework for SOIL d) – (true) linework for Yeovil soil parent material map.

It is clear from the sum of the difference between the actual parent material extent and the various predicted extents, that a grid based systematic sample provides a more accurate understanding of the parent materials present than the same number of points collected in a traditional, clustered survey pattern (8% vs. 20%, see Table 39). The difference between the truth and the prediction based on the national composition of the SOIL units was particularly poor (87% difference). This arose mostly because of the over prediction of a unit which was not actually in the study area (FqF1 – thin drift, passing to sand or soft sandstone). According to SOIL, this unit makes up 28.7% of the area. There is confusion with FyF1 (thin drift, passing to soft shale or siltstone) which the reference map shows to cover 29% of the area. There is similarity between these units, yet the difference in texture can give rise to quite different hydrological characteristics. On the basis of this analysis, it is suggested that the National Soil Map (SOIL) is not used as the sole predictor of which parent material classes are likely to be present in the area.

As the grid-based sample produced the result with the lowest difference in the extents from the reference map, sample grid densities from 60 m (65,000 points) to 3000 m (9 points) were compared (Figure 54). Once more, the sum of the difference was calculated for each map unit, and this is plotted for each sample density in Figure 54.

Figure 54 shows how well grid samples of decreasing sample density predict parent material composition and extents. More closely spaced sampling regimes perform better, but are more costly, so a value judgement needs to be made with regards to the level of accuracy required.

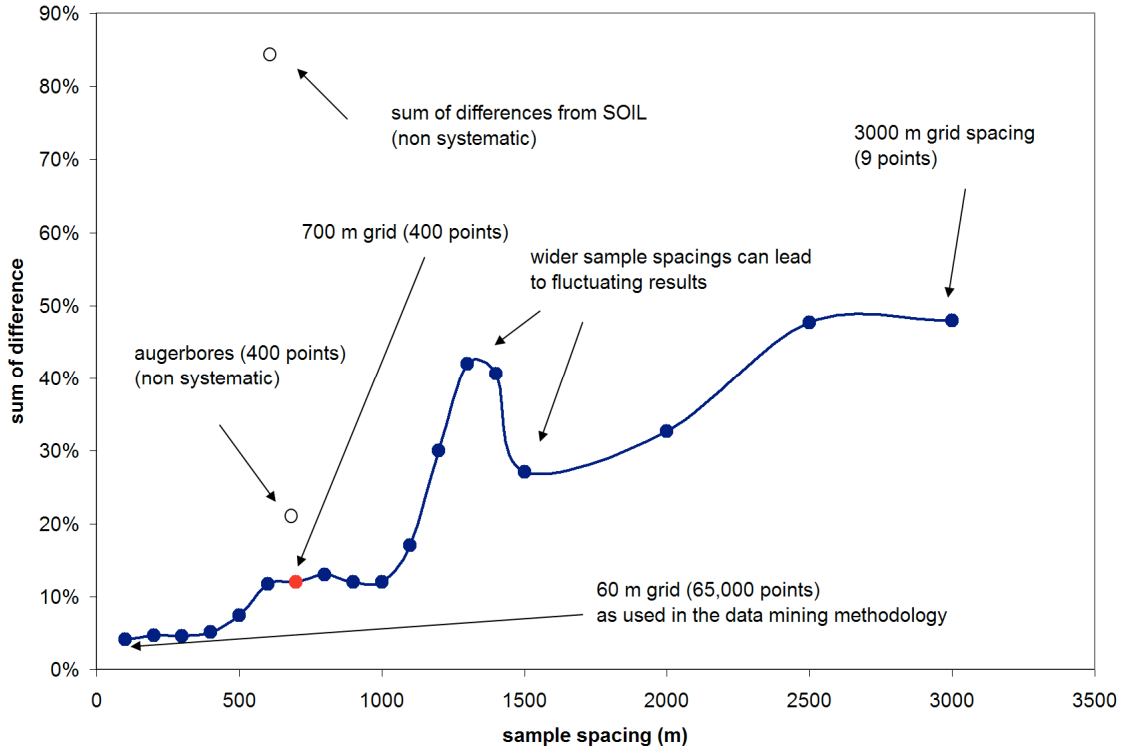


Figure 54 –The sum of differences in predicted parent material extents compared with the actual parent material extent, with increasing sample spacing
Note: Grids with smaller spacing tend to also identify more parent material units. A sample spacing of greater than 1000 m tends to produce less accurate predictions of parent material extent. They are also more volatile due to the effect of single sample points on the predicted values.

8.4.4 Testing evidence layer association (Figure 52d)

In the expert knowledge and data mining methodologies, SLOPE did not produce maps of as high a value as did the GEOLOGY and SOIL evidence layers. In typical data mining scenarios, a large number of datasets might be prepared, and, in a stepwise manner, added or removed until the optimal result is achieved. This process was used in the expert knowledge and data mining methodologies. This stepwise approach can be time consuming (Stockwell, 2006) since it involves many model runs. But more important in this research is the time required to extract and formalise expert knowledge. Therefore, it is useful to investigate the relationship between potential evidence layers and parent material before the laborious task of formalising expert knowledge is undertaken. For this purpose, two approaches were used.

8.4.4.1 Pairwise association – Chi squared

With Pearson's chi-squared (Plackett, 1983), it was possible to test for pairwise association between predicting evidence layers and soil parent material, based on a sparse (700 m) field sample. Statistical significance tests generate p-values to evaluate evidence against null hypotheses. In this analysis, the p-values presented in Table 1 were calculated to test the null hypotheses of no association between parent material and various evidence layers. Small values, such as those seen for SOIL and GEOLOGY evidence layers, do not necessarily reflect the usefulness of the layers to predict parent material, but do reflect evidence for an association between these layers and parent material. Conversely, the higher p-values obtained with SLOPE indicate less evidence of association, and less likelihood that these evidence layers would add useful detail to the model.

Table 40 – Comparison of p-value results from Pearson's chi-squared test.
Note: chi-squared tests for evidence of association between the predicting layer and parent material. Small values indicate association.

Study Area	SOIL	GEOLOGY	Slope
Worksop	1.3×10^{-44}	1.1×10^{-14}	0.26
Needwood Forest	2.0×10^{-86}	5.2×10^{-36}	0.0006
Yeovil	2.5×10^{-54}	7.0×10^{-69}	0.04

Often, a 5% significance level is adopted as the criterion to reject or not reject the null hypothesis. No such formal test is used here, but rather the chi squared test is used to provide a quantitative assessment of which evidence layers demonstrate strong evidence for associations with parent material, and therefore which may benefit from the extraction of expert knowledge.

Because chi squared tests examine the association between a pair of datasets, they are particularly suited for use in models similar to Expector, as such models assume conditional independence. Nevertheless, in some methods, it is not ideal if layers are excluded exclusively on a basis of low pairwise association. As in multivariate data mining exercises (those used in this research have been bivariate), it is often better to do a stepwise removal or addition of layers, until the best model result is obtained. An alternative approach to this is to use multinomial logistic regression to identify evidence layers of potential use.

8.4.4.2 Multinomial logistic regression

Multinomial logistic regression is a means to provide a sense of the relative explanatory potential of the evidence layers. The improvement in deviance (a measure of goodness of fit) over a simple model with prediction by the prior probabilities of the hypotheses (parent material classes) is considered, when various evidence layers are used as

explanatory variables for parent material. For the Yeovil area (the most complex), the improvement in deviance using all three evidence layers was 106634.

The improvement on the simple model for the evidence layers individually can be expressed as a fraction of this total improvement. These improvements are displayed in Table 41 Also shown in Table 41 is the percentage of improvement gained when a third evidence layer is added to the other two, representing unique information provided by that evidence layer.

Table 41 - Deviances of the individual evidence layers

Evidence Layer	Deviance	Improvement Individually	Unique Improvement
GEOLOGY	89999	84.4%	23.6%
SOIL	76456	71.0%	13.0%
SLOPE	11303	10.0%	1.6%

Given the size of the data sets, any additional source of information is likely to generate a statistically significant improvement. However, the practical importance is shown in the percentage improvements, and it is quite clear from these analyses that the dominant explanatory variables are GEOLOGY and SOIL. Therefore, for this fourth methodology model inputs were created for GEOLOGY and SOIL evidence layers.

Additional potentially useful evidence layers (e.g. other DTM derivatives or classified remote sensing data) could be investigated with the statistical methods used here. Then appropriate expert knowledge could be obtained through structured interviews, assuming experts are available. However, additional evidence layers have not been investigated in this methodology so as to maintain consistency of predictive information with previous methods.

8.4.5 The preparation of model inputs (Figure 52e)

The same procedure for generating model inputs for expert knowledge (see section 6.4) was used in this methodology. The lack of quantitative data on the relationships between the evidence layer classes and parent material classes was shown to be a hindrance to the expert knowledge methodology. Therefore, in this combined approach, additional quantitative data from the sparse samples was used to amend the model inputs derived from expert knowledge.

In order to create these new inputs, two confusion matrices showing relationships between the evidence layers and the soil parent material were displayed side-by side (Figure 55). One was derived from the expert knowledge (in green), the other from the data mining of the sparse field samples (in red). Aspects of both matrices were used to create a unified input.

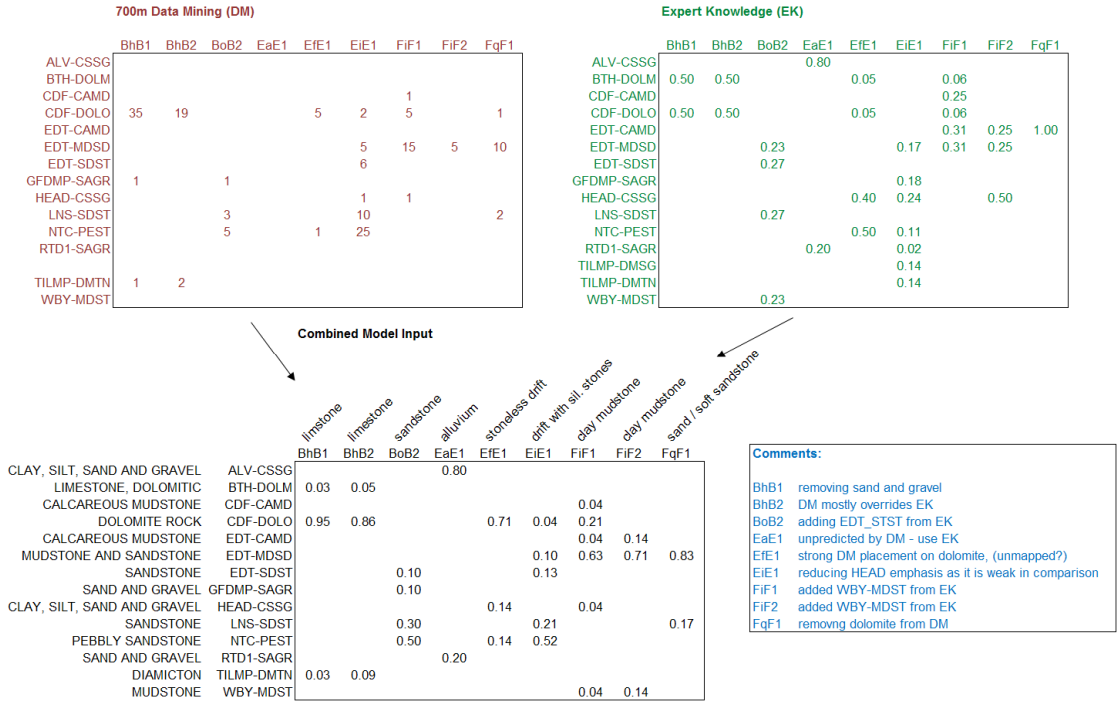


Figure 55 – The process of combining expert knowledge and 700 m data mining inputs
Note: not all classes are predicted by the data mining method (DM) (red) and the Expert Knowledge (EK) (green) can have poorly quantified predictions. Combining aspects of both (as described in the blue comment box) leads to the combined model input. This figure shows the generation of the GEOLOGY input for Worksop.

As an example of the thought process undertaken for this process of combination, the combined input for BhB1 (limestone) in Figure 55 will be considered. The red data mining sample had 37 points with BhB1. Over 95% of these were found on the GEOLOGY unit CDF-DOLO (dolomite). One point was found on TILMP-DMTN (till) and one on GFDMP-SAGR (sand and gravel). Considering the green expert knowledge input, it was seen that the literature indicates that this BhB1 unit can be found on either of the dolomitic units in the area (BTH-DOLM or CDF-DOLM) but little was known from the expert knowledge about the dominance of the CDF-DOLM unit. The sand and gravel point was discarded by the expert knowledge, but the diamicton was retained due to possible glacial reworking of the limestone. The dominance seen from the data mining sample (35 points on CDF-DOLO) was retained, and the combined input assigned a probability of 0.95 to CDF-DOLO and 0.03 to both BTH-DOLM and TILMP-DMTN.

Because the 700 m data mining confusion matrix (top right, Figure 55) had relatively few points, occasionally this sample would miss geological and parent material units which were known from expert knowledge (top left, Figure 55) and the evidence layers themselves, to be present in the study areas. This was increasingly common for the very sparse (2100 m and 2800 m samples). In these cases, expert knowledge was more strongly relied upon to guide the model input. In other cases, the expert knowledge information was refined by the data mined relationships where units were consistently predicted over different evidence classes, or in different proportions to those expected by expert knowledge.

8.4.6 Model runs (Figure 52f)

Once the combined model inputs were created, these were entered into the probability model. If a real field sample was undertaken, the results of the point specific sampling could be used to provide a quantitative assessment of the accuracy of the evidence layers in the map purity tables (see section 4.3). As no real field survey sampling was carried out in this research, the models were run with an estimated map purity of 95%,

as in the previous methodologies. Models were run using GEOLOGY and SOIL evidence layers, both individually and together.

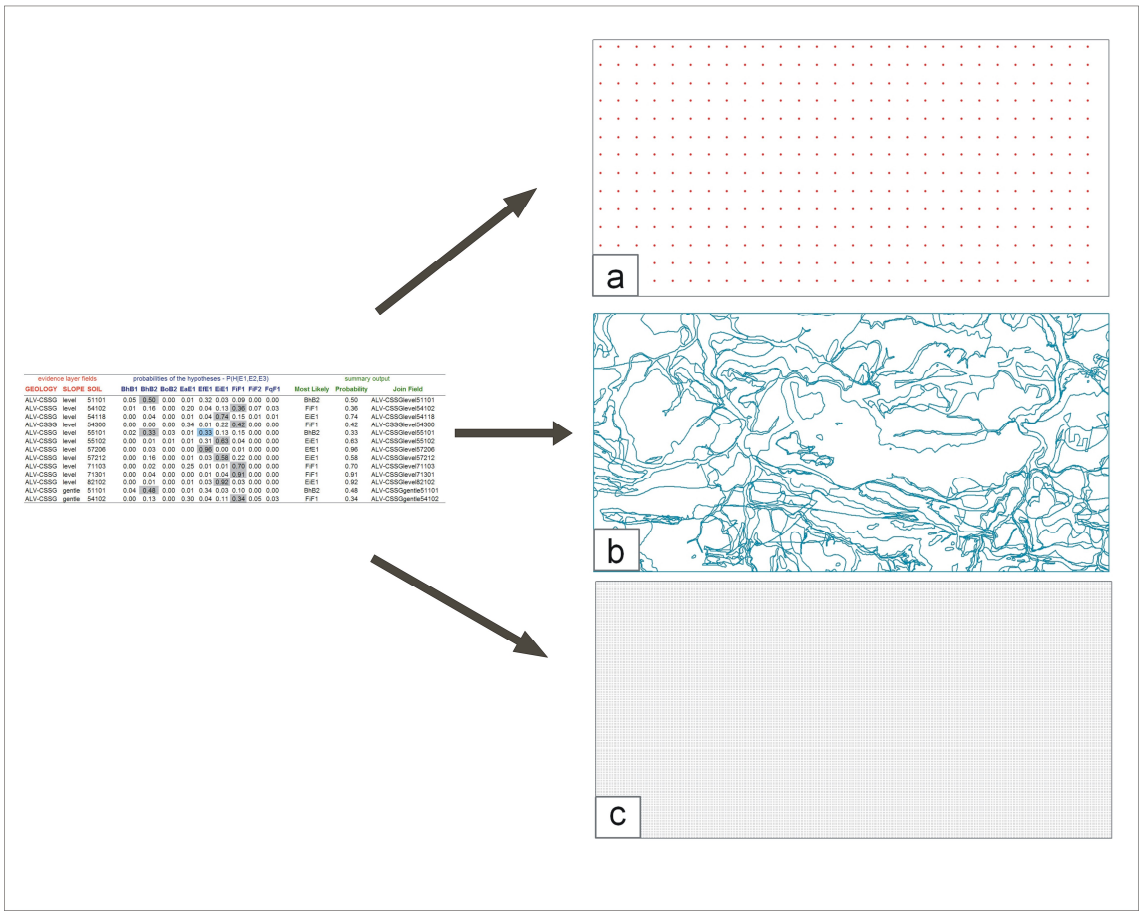
It was hypothesised that the addition of expert knowledge to the sparse sample date will produce maps of higher value than could be obtained just using the sparse sample. To test this, models were also run based entirely on the data from the sparse sample, with no additional expert knowledge.

8.4.7 Success assessments and class amalgamations (Figure 52g)

The results of the models were attached to the shapefiles of the field data sample (Figure 56a), and assessed in the same manner as in previous methodologies. Potential class amalgamations were identified and these were tested against the reference of the field sample. The class amalgamations leading to the highest map values (ψ_3) were noted.

8.4.8 The creation of the final maps (Figure 52h)

To reduce file size and increase GIS functionality and aesthetic quality, the final maps were created using the original vector linework of the input layers (Figure 56b) rather than the 60 m grid point shapefile used in the previous methods. The full model results were attributed to the polygons allowing not only the display of the most likely parent material, but additionally, the probability of each of the parent materials (Table 9, p70). This has the advantage of allowing any future models to make use of the full detail of the probability model output.



evidence layer fields			probabilities of the hypotheses - P(H|E1,E2,E3)								summary output		
GEOLOGY	SLOPE	SOIL	BhB1	BhB2	BoB2	EaE1	EfE1	EiE1	FiF1	FiF2	FoF1	Most Likely	Probability	Join Field
ALV-CSSG	level	51101	0.05	0.50	0.00	0.01	0.32	0.03	0.09	0.00	0.00	BhB2	0.50	ALV-CSSGlevel51101
ALV-CSSG	level	54102	0.01	0.16	0.00	0.20	0.04	0.13	0.36	0.07	0.03	FiF1	0.36	ALV-CSSGlevel54102
ALV-CSSG	level	54118	0.00	0.04	0.00	0.01	0.04	0.74	0.15	0.01	0.01	EiE1	0.74	ALV-CSSGlevel54118
ALV-CSSG	level	04300	0.00	0.00	0.00	0.34	0.01	0.22	0.42	0.00	0.00	FiF1	0.42	ALV-CSSGlevel04300
ALV-CSSG	level	55101	0.02	0.33	0.03	0.01	0.33	0.13	0.15	0.00	0.00	BhB2	0.33	ALV-CSSGlevel55101
ALV-CSSG	level	55102	0.00	0.01	0.01	0.01	0.31	0.63	0.04	0.00	0.00	EiE1	0.63	ALV-CSSGlevel55102
ALV-CSSG	level	57206	0.00	0.03	0.00	0.00	0.96	0.00	0.01	0.00	0.00	EiE1	0.96	ALV-CSSGlevel57206
ALV-CSSG	level	57212	0.00	0.16	0.00	0.01	0.03	0.58	0.22	0.00	0.00	EiE1	0.58	ALV-CSSGlevel57212
ALV-CSSG	level	71103	0.00	0.02	0.00	0.25	0.01	0.01	0.70	0.00	0.00	FiF1	0.70	ALV-CSSGlevel71103
ALV-CSSG	level	71301	0.00	0.04	0.00	0.00	0.01	0.04	0.91	0.00	0.00	FiF1	0.91	ALV-CSSGlevel71301
ALV-CSSG	level	82102	0.00	0.01	0.00	0.01	0.03	0.92	0.03	0.00	0.00	EiE1	0.92	ALV-CSSGlevel82102
ALV-CSSG	gentle	51101	0.04	0.48	0.00	0.01	0.34	0.03	0.10	0.00	0.00	BhB2	0.48	ALV-CSSGgentle51101
ALV-CSSG	gentle	54102	0.00	0.13	0.00	0.30	0.04	0.11	0.34	0.05	0.03	FiF1	0.34	ALV-CSSGgentle54102

Figure 56 - Three shapefiles to which model outputs are joined for analysis
Note: Model output (see Table 9) is joined to three GIS shapefiles: a) 700 m sparse sample used for amalgamation analyses; b) linework for cartographic display c) 60 m grid used for comparison with previous methodologies

In a real world situation, the maps would be based on the existing linework (Figure 56b) and the map and class values would be calculated from the sparse sampling (Figure 56a) or testing in a known area.

8.4.9 Analysis for comparison with previous methods

A full analysis of the relative success of the models applied to the detailed 60 m grid (Figure 56c) was compiled for consistent comparison with previous methods. It is from these maps that the reported results were calculated. The map values achieved using the detailed 60 m sample were consistently marginally lower than those based on the

sparser sample in section 8.4.7 (Table 42). This effect was most noticeable in the map value (ψ_3) assessment.

Table 42 – Comparison of the results from different testing densities (based on test N75)
Note: This table shows the results of the same model and inputs, assessed on two different testing grids.

	ψ_3	θ_1	κ
700 m testing grid	1.78	0.67	0.48
60 m testing grid	1.50	0.66	0.47

Thus, it must be remembered that when judgements are based on a smaller number of sample points, the resulting map across the whole area is unlikely to achieve the map value indicated by the initial test. This will be particularly important if the test points were also used in the creation of the model inputs, as they were in this case.

8.5 Combined methodology results

For consistency with previous methods, the results for this methodology are presented based on the full 60 m grid analysis. The results from the sparse data sample for the data mining inputs, as well as the combined approach using both expert knowledge and the increasingly sparse samples are displayed in Table 43 to Table 48.

Table 43 – Different sample density data mining results for Worksop

Note: Presenting results for parent material maps created using the sample data input only, with no expert knowledge. The 60 m samples are repeated from the data mining methodology. For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Worksop (data mining)

Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
GEOLOGY, SOIL 60m	0.55	**0.65**	1.60	9	8	7	5	4	1	-	1	0%	W41
GEOLOGY, SOIL 700m	0.54	**0.64**	1.45	9	8	7	4	3	1	-	1	0%	W51
GEOLOGY, SOIL 1400m	0.51	**0.62**	1.27	9	6	5	4	3	1	-	1	9%	W52
GEOLOGY, SOIL 2100m	0.49	**0.61**	1.04	9	5	4	3	3	1	-	1	22%	W53
GEOLOGY, SOIL 2800m	0.48	**0.60**	1.05	9	4	3	3	3	1	-	1	29%	W54
GEOLOGY 60m	0.47	**0.59**	0.94	9	6	3	3	3	1	-	1	12%	W38
GEOLOGY 700m	0.46	**0.59**	0.89	9	5	3	3	3	1	-	1	13%	W55
GEOLOGY 1400m	0.38	**0.51**	0.64	9	4	3	3	3	-	-	1	29%	W56
GEOLOGY 2100m	0.45	**0.58**	0.83	9	3	3	3	3	1	-	1	33%	W57
GEOLOGY 2800m	0.45	**0.58**	0.83	9	3	3	3	3	1	-	1	33%	W58
SOIL 60m	0.54	**0.64**	1.49	9	6	6	4	4	1	-	1	22%	W43
SOIL 700m	0.54	**0.64**	1.47	9	7	6	4	4	1	-	1	7%	W59
SOIL 1400m	0.51	**0.61**	1.18	9	4	4	3	3	1	-	1	29%	W60
SOIL 2100m	0.45	**0.57**	0.87	9	3	3	3	3	1	-	1	33%	W61
SOIL 2800m	0.45	**0.62**	0.64	9	3	2	2	2	1	1	2	29%	W62
GEOLOGY, SOIL 60m (A)	0.76	**0.83**	2.17	7	6	6	5	5	2	2	2	0%	W48
GEOLOGY, SOIL 700m (A)	0.74	**0.81**	1.99	7	6	6	5	4	2	2	2	0%	W63
GEOLOGY, SOIL 1400m (A)	0.72	**0.80**	1.80	8	5	5	5	4	2	2	2	2%	W64
GEOLOGY, SOIL 2100m (A)	0.84	**0.89**	1.44	6	3	3	3	3	3	3	3	2%	W65
GEOLOGY, SOIL 2800m (A)	0.75	**0.82**	1.71	7	4	4	4	4	2	2	3	2%	W66
GEOLOGY 60m (A)	0.70	**0.80**	1.27	6	6	3	3	3	2	2	3	0%	W45
GEOLOGY 700m (A)	0.67	**0.78**	1.22	7	4	3	3	3	2	2	2	6%	W67
GEOLOGY 1400m (A)	0.75	**0.84**	1.18	6	3	3	3	3	2	3	3	2%	W68
GEOLOGY 2100m (A)	0.75	**0.84**	1.18	6	3	3	3	3	2	3	3	2%	W69
GEOLOGY 2800m (A)	0.75	**0.84**	1.18	6	3	3	3	3	2	3	3	2%	W70
SOIL 60m (A)	0.77	**0.84**	2.22	7	7	6	5	5	2	2	2	0%	W50
SOIL 700m (A)	0.77	**0.84**	2.22	7	6	6	5	5	2	2	2	0%	W71
SOIL 1400m (A)	0.77	**0.84**	1.85	7	4	4	4	4	2	2	3	2%	W72
SOIL 2100m (A)	0.77	**0.85**	1.36	6	3	3	3	3	2	2	4	2%	W73
SOIL 2800m (A)	0.75	**0.84**	1.15	5	3	3	3	3	2	3	3	2%	W74

Table 44 - Combined methodology results for Worksop

Note: Presenting results for parent material maps created using the combined methodology. For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Worksop (combined)

Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
GEOLOGY, SOIL 700m	0.55	**0.65**	1.53	9	9	5	5	4	1	-	1	0%	W75
GEOLOGY, SOIL 1400m	0.51	**0.61**	1.17	9	8	5	3	3	1	-	1	17%	W76
GEOLOGY, SOIL 2100m	0.49	**0.61**	1.02	9	8	4	3	3	1	-	1	15%	W77
GEOLOGY, SOIL 2800m	0.50	**0.61**	1.14	9	8	4	3	3	1	-	1	17%	W78
GEOLOGY 700m	0.45	**0.58**	0.88	9	7	3	3	3	1	-	1	11%	W79
GEOLOGY 1400m	0.35	**0.48**	0.58	9	6	3	3	3	-	-	1	26%	W80
GEOLOGY 2100m	0.35	**0.48**	0.58	9	5	3	3	3	-	-	1	28%	W81
GEOLOGY 2800m	0.35	**0.48**	0.58	9	5	3	3	3	-	-	1	28%	W82
SOIL 700m	0.53	**0.63**	1.47	9	7	6	4	4	1	-	1	22%	W83
SOIL 1400m	0.51	**0.62**	1.20	9	5	4	3	3	1	-	1	29%	W84
SOIL 2100m	0.51	**0.62**	1.16	9	4	4	4	3	1	-	1	29%	W85
SOIL 2800m	0.51	**0.61**	1.18	9	5	4	3	3	1	-	1	29%	W86
GEOLOGY, SOIL 700m (A)	0.73	**0.81**	1.95	7	7	5	5	4	2	2	2	0%	W87
GEOLOGY, SOIL 1400m (A)	0.71	**0.80**	1.66	8	7	4	4	4	2	2	2	2%	W88
GEOLOGY, SOIL 2100m (A)	0.75	**0.83**	1.43	7	6	3	3	3	3	2	3	0%	W89
GEOLOGY, SOIL 2800m (A)	0.73	**0.81**	1.65	7	6	4	4	4	2	2	3	2%	W90
GEOLOGY 700m (A)	0.70	**0.80**	1.27	6	6	3	3	3	2	2	3	0%	W91
GEOLOGY 1400m (A)	0.76	**0.84**	1.23	6	5	3	3	3	2	3	3	0%	W92
GEOLOGY 2100m (A)	0.76	**0.84**	1.22	6	4	3	3	3	2	3	3	2%	W93
GEOLOGY 2800m (A)	0.76	**0.84**	1.22	6	4	3	3	3	2	3	3	2%	W94
SOIL 700m (A)	0.77	**0.84**	2.22	7	7	6	5	5	2	2	2	0%	W95
SOIL 1400m (A)	0.78	**0.84**	1.87	6	5	4	4	4	2	2	3	2%	W96
SOIL 2100m (A)	0.80	**0.86**	1.74	6	4	4	4	3	3	2	3	2%	W97
SOIL 2800m (A)	0.77	**0.84**	1.84	6	5	4	4	4	2	2	3	2%	W98

Table 45 - Different sample density data mining results for Needwood Forest

Note: Presenting results for parent material maps created using the sample data input only, with no expert knowledge. The 60 m samples are repeated from the data mining methodology. For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Needwood Forest (data mining) Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
GEOLOGY, SOIL 60m	0.42	**0.62**	**1.49**	11	8	8	5	4	-	-	1	7%	N41
GEOLOGY, SOIL 700m	0.45	**0.62**	**1.43**	11	7	6	5	4	1	-	1	4%	N51
GEOLOGY, SOIL 1400m	0.41	**0.60**	**1.10**	11	7	6	4	3	-	-	1	2%	N52
GEOLOGY, SOIL 2100m	0.30	**0.57**	**0.74**	11	5	4	4	3	-	-	1	19%	N53
GEOLOGY, SOIL 2800m	-0.12	**0.28**	**0.11**	11	5	2	2	-	-	-	1	19%	N54
GEOLOGY 60m	0.19	**0.60**	**0.94**	11	6	5	3	3	-	-	1	14%	N38
GEOLOGY 700m	0.15	**0.56**	**0.86**	11	5	4	3	3	-	-	1	23%	N55
GEOLOGY 1400m	0.13	**0.56**	**0.56**	11	4	2	2	2	-	-	1	23%	N56
GEOLOGY 2100m	0.16	**0.57**	**0.60**	11	3	3	2	2	-	-	1	26%	N57
GEOLOGY 2800m	0.01	**0.52**	**0.28**	11	2	1	1	1	-	-	1	34%	N58
SOIL 60m	0.40	**0.63**	**1.14**	11	5	5	5	4	-	-	1	14%	N43
SOIL 700m	0.44	**0.65**	**1.00**	11	5	4	4	3	-	-	1	11%	N59
SOIL 1400m	0.41	**0.61**	**0.96**	11	6	5	5	3	-	-	1	3%	N60
SOIL 2100m	0.25	**0.54**	**0.56**	11	4	4	3	2	-	-	1	26%	N61
SOIL 2800m	-0.08	**0.21**	**0.04**	11	5	1	1	-	-	-	1	7%	N62
GEOLOGY, SOIL 60m (A)	0.51	**0.69**	**1.61**	9	7	7	5	5	-	2	2	3%	N48
GEOLOGY, SOIL 700m (A)	0.48	**0.65**	**1.52**	11	7	6	5	5	1	1	2	1%	N63
GEOLOGY, SOIL 1400m (A)	0.54	**0.85**	**1.21**	9	4	4	4	3	1	2	3	2%	N64
GEOLOGY, SOIL 2100m (A)	0.44	**0.72**	**0.91**	10	4	4	4	4	1	2	2	3%	N65
GEOLOGY, SOIL 2800m (A)	0.39	**0.94**	**0.56**	7	2	2	2	2	1	2	5	1%	N66
GEOLOGY 60m (A)	0.50	**0.91**	**1.21**	7	4	4	3	3	1	1	5	3%	N45
GEOLOGY 700m (A)	0.44	**0.90**	**1.24**	9	4	3	3	3	1	1	4	4%	N67
GEOLOGY 1400m (A)	0.58	**0.94**	**0.82**	7	2	2	2	2	1	1	5	3%	N68
GEOLOGY 2100m (A)	0.58	**0.95**	**0.82**	7	2	2	2	2	1	1	5	3%	N69
GEOLOGY 2800m (A)	0.03	**0.83**	**0.32**	7	2	1	1	1	1	1	5	3%	N70
SOIL 60m (A)	0.54	**0.75**	**1.18**	9	4	4	4	4	1	2	2	7%	N50
SOIL 700m (A)	0.56	**0.75**	**1.02**	11	5	4	4	3	1	1	3	1%	N71
SOIL 1400m (A)	0.47	**0.65**	**1.10**	11	5	5	5	4	-	1	2	3%	N72
SOIL 2100m (A)	0.40	**0.70**	**0.76**	10	4	4	3	3	-	2	2	3%	N73
SOIL 2800m (A)	-0.06	**0.76**	**0.29**	8	2	1	1	1	1	1	4	7%	N74

Table 46 - Combined methodology results for Needwood Forest

Note: Presenting results for parent material maps created using the combined methodology. For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Needwood Forest (combined)													
Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
GEOLOGY, SOIL 700m	0.47	**0.66**	1.50	11	8	6	5	5	-	-	1	10%	N75
GEOLOGY, SOIL 1400m	0.44	**0.62**	1.45	11	9	7	5	5	-	-	1	0%	N76
GEOLOGY, SOIL 2100m	0.39	**0.62**	1.24	11	8	6	5	5	-	-	1	1%	N77
GEOLOGY, SOIL 2800m	0.36	**0.53**	1.07	11	9	6	5	3	-	-	1	2%	N78
GEOLOGY 700m	0.16	**0.56**	0.87	11	6	4	3	3	-	-	1	23%	N79
GEOLOGY 1400m	0.12	**0.55**	0.56	11	6	2	2	2	-	-	1	21%	N80
GEOLOGY 2100m	0.14	**0.57**	0.82	11	5	4	3	3	-	-	1	13%	N81
GEOLOGY 2800m	0.16	**0.56**	0.73	11	5	4	3	3	-	-	1	13%	N82
SOIL 700m	0.47	**0.67**	1.42	11	7	5	5	5	-	-	1	11%	N83
SOIL 1400m	0.34	**0.56**	0.93	11	7	6	5	2	-	-	1	3%	N84
SOIL 2100m	0.35	**0.60**	1.02	11	6	6	4	4	-	-	1	6%	N85
SOIL 2800m	0.20	**0.52**	0.66	11	6	4	3	2	-	-	1	13%	N86
GEOLOGY, SOIL 700m (A)	0.54	**0.73**	1.56	10	8	6	5	5	1	1	2	4%	N87
GEOLOGY, SOIL 1400m (A)	0.61	**0.75**	1.84	10	7	6	6	6	1	2	2	0%	N88
GEOLOGY, SOIL 2100m (A)	0.42	**0.64**	1.32	11	7	6	6	5	-	1	2	1%	N89
GEOLOGY, SOIL 2800m (A)	0.39	**0.57**	1.23	10	7	7	6	5	-	2	2	2%	N90
GEOLOGY 700m (A)	0.46	**0.90**	1.25	8	5	3	3	3	1	1	4	4%	N91
GEOLOGY 1400m (A)	0.55	**0.93**	0.87	6	4	3	2	2	1	2	5	0%	N92
GEOLOGY 2100m (A)	0.20	**0.72**	0.93	10	4	4	3	3	1	1	3	9%	N93
GEOLOGY 2800m (A)	0.60	**0.94**	1.06	7	3	3	3	3	1	1	5	2%	N94
SOIL 700m (A)	0.62	**0.80**	1.42	9	6	4	4	4	2	2	2	4%	N95
SOIL 1400m (A)	0.54	**0.72**	1.36	10	5	5	5	4	-	2	2	3%	N96
SOIL 2100m (A)	0.37	**0.62**	1.07	11	6	6	5	4	-	1	2	3%	N97
SOIL 2800m (A)	0.30	**0.93**	1.09	7	4	4	4	2	1	2	5	2%	N98

Table 47 – Different sample density data mining results for Yeovil

Note: Presenting results for parent material maps created using the sample data input only, with no expert knowledge. The 60 m samples are repeated from the data mining methodology. For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Yeovil (data mining) Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ> 0.2	C ξ> 0.4	C ξ> 0.5	C ξ> 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
GEOLOGY, SOIL 60m	0.58	**0.66**	2.42	17	12	10	8	6	-	-	1	6%	Y41
GEOLOGY, SOIL 700m	0.56	**0.64**	1.77	17	12	8	7	4	-	-	1	1%	Y51
GEOLOGY, SOIL 1400m	0.50	**0.59**	1.43	17	11	7	5	5	-	-	1	2%	Y52
GEOLOGY, SOIL 2100m	0.44	**0.55**	0.98	17	7	5	5	3	-	-	1	10%	Y53
GEOLOGY, SOIL 2800m	0.49	**0.59**	1.27	17	7	6	6	4	-	-	1	10%	Y54
GEOLOGY 60m	0.55	**0.64**	2.13	17	9	9	7	6	-	-	1	9%	Y38
GEOLOGY 700m	0.54	**0.62**	1.65	17	8	8	6	4	-	-	1	9%	Y55
GEOLOGY 1400m	0.46	**0.56**	1.26	17	7	6	6	5	-	-	1	20%	Y56
GEOLOGY 2100m	0.30	**0.46**	0.48	17	5	3	3	2	-	-	1	26%	Y57
GEOLOGY 2800m	0.42	**0.54**	1.00	17	7	6	4	3	-	-	1	10%	Y58
SOIL 60m	0.51	**0.62**	1.37	17	7	7	6	3	-	-	1	12%	Y43
SOIL 700m	0.51	**0.61**	1.39	17	9	8	6	3	-	-	1	5%	Y59
SOIL 1400m	0.47	**0.58**	1.08	17	8	6	4	3	-	-	1	8%	Y60
SOIL 2100m	0.42	**0.54**	0.81	17	6	5	3	3	-	-	1	14%	Y61
SOIL 2800m	0.44	**0.56**	0.93	17	6	5	4	3	-	-	1	14%	Y62
GEOLOGY, SOIL 60m (A)	0.58	**0.66**	2.42	16	12	10	8	6	-	-	1	6%	Y48
GEOLOGY, SOIL 700m (A)	0.56	**0.64**	1.77	16	12	8	7	4	-	-	1	1%	Y63
GEOLOGY, SOIL 1400m (A)	0.50	**0.59**	1.43	17	11	7	5	5	-	-	1	2%	Y64
GEOLOGY, SOIL 2100m (A)	0.49	**0.59**	1.15	16	6	6	6	4	-	1	2	10%	Y65
GEOLOGY, SOIL 2800m (A)	0.57	**0.66**	1.39	13	6	6	6	5	-	2	4	4%	Y66
GEOLOGY 60m (A)	0.58	**0.66**	2.15	15	9	9	7	6	-	1	2	6%	Y45
GEOLOGY 700m (A)	0.57	**0.65**	1.71	15	7	7	6	5	-	1	2	9%	Y67
GEOLOGY 1400m (A)	0.54	**0.62**	1.56	16	7	6	6	5	-	-	1	10%	Y68
GEOLOGY 2100m (A)	0.43	**0.61**	0.69	13	4	4	4	2	-	2	3	10%	Y69
GEOLOGY 2800m (A)	0.47	**0.59**	1.11	14	6	6	6	3	-	2	2	6%	Y70
SOIL 60m (A)	0.51	**0.62**	1.37	16	7	7	6	3	-	-	1	12%	Y50
SOIL 700m (A)	0.51	**0.61**	1.39	17	9	8	6	3	-	-	1	5%	Y71
SOIL 1400m (A)	0.47	**0.58**	1.08	17	8	6	4	3	-	-	1	8%	Y72
SOIL 2100m (A)	0.44	**0.56**	0.90	16	6	6	4	3	-	1	2	10%	Y73
SOIL 2800m (A)	0.58	**0.69**	1.07	13	4	4	4	3	-	1	5	11%	Y74

Table 48 - Combined methodology results for Yeovil

Note: Presenting results for parent material maps created using the combined methodology. For explanations of the headings, see Table 9, page 70. (A) indicates tests with class amalgamation.

Yeovil (combined) Method	κ	θ1	ψ3	Total Classes	Effective Classes	C ξ > 0.2	C ξ > 0.4	C ξ > 0.5	C ξ > 0.8	Amalg. Classes	Max. Class Size	% Unpredictable	Test
GEOLOGY, SOIL 700m	0.57	**0.65**	1.91	17	14	9	7	4	-	-	1	1%	Y75
GEOLOGY, SOIL 1400m	0.56	**0.64**	1.77	17	12	8	7	4	-	-	1	1%	Y76
GEOLOGY, SOIL 2100m	0.45	**0.55**	1.35	17	13	7	6	5	-	-	1	4%	Y77
GEOLOGY, SOIL 2800m	0.46	**0.56**	1.46	17	12	8	6	4	-	-	1	4%	Y78
GEOLOGY 700m	0.55	**0.63**	1.97	17	9	8	7	6	-	-	1	9%	Y79
GEOLOGY 1400m	0.53	**0.62**	1.79	17	9	7	7	6	-	-	1	9%	Y80
GEOLOGY 2100m	0.50	**0.59**	1.65	17	9	7	7	6	-	-	1	9%	Y81
GEOLOGY 2800m	0.47	**0.57**	1.42	17	9	7	6	4	-	-	1	6%	Y82
SOIL 700m	0.51	**0.61**	1.39	17	9	8	6	3	-	-	1	5%	Y83
SOIL 1400m	0.48	**0.59**	1.07	17	8	6	3	3	-	-	1	8%	Y84
SOIL 2100m	0.44	**0.56**	0.89	17	7	6	3	3	-	-	1	10%	Y85
SOIL 2800m	0.47	**0.58**	1.08	17	7	7	4	3	-	-	1	10%	Y86
GEOLOGY, SOIL 700m (A)	0.59	**0.67**	1.93	15	13	8	7	5	-	1	2	1%	Y87
GEOLOGY, SOIL 1400m (A)	0.56	**0.64**	1.77	16	12	8	7	4	-	-	1	1%	Y88
GEOLOGY, SOIL 2100m (A)	0.53	**0.62**	1.61	13	10	7	7	6	-	3	3	0%	Y89
GEOLOGY, SOIL 2800m (A)	0.55	**0.64**	1.61	13	8	7	6	6	-	3	3	3%	Y90
GEOLOGY 700m (A)	0.58	**0.66**	1.98	15	9	8	7	6	-	1	2	6%	Y91
GEOLOGY 1400m (A)	0.57	**0.66**	1.79	14	8	7	7	6	-	2	2	6%	Y92
GEOLOGY 2100m (A)	0.55	**0.64**	1.67	14	8	7	7	6	-	2	2	6%	Y93
GEOLOGY 2800m (A)	0.55	**0.64**	1.57	12	7	6	6	6	-	2	4	4%	Y94
SOIL 700m (A)	0.51	**0.61**	1.39	16	9	8	6	3	-	-	1	5%	Y95
SOIL 1400m (A)	0.49	**0.60**	1.08	15	8	6	4	3	-	1	2	4%	Y96
SOIL 2100m (A)	0.47	**0.59**	1.02	16	6	6	4	4	-	1	2	10%	Y97
SOIL 2800m (A)	0.47	**0.58**	1.08	17	7	7	4	3	-	-	1	10%	Y98

8.6 Discussion

This discussion will consider the results of the combined method with those of the previous methodologies. Figure 57 to Figure 59 display the relative successes of the expert knowledge (grey lines), data mining using the different sample spacings (dark blue lines), and combined (light blue lines) methodologies. These figures provide a useful basis for discussions around the success of the different methodologies in the different areas. They include brief comments, illustrative of key points.

Across all study areas, a clear drop off in the map values achieved by data mining can be seen as the sample spacing for the model input increases. In the Yeovil area, a quick drop in map value is seen when comparing the whole area (60 m sample) to a more achievable 700 m sample (Figure 59). However, in the Needwood Forest and Worksop areas, the decrease in map value for models trained on 60 m and 700 m samples is more gentle. This difference appears to be predominantly due to the relative complexity of the Yeovil area, which contains 17 parent material units, compared to the 11 and 9 units of Needwood Forest and Worksop. The maps resulting from the models using data mining inputs in Needwood Forest are shown in Figure 60 to Figure 63. Over these four maps, with increasing sample spacing 700 m to 2800 m, the increase in incorrect prediction (red on maps 'c') becomes apparent. Additionally, less extensive units such as DbD1 (non-calcareous gravel) and AaA3 (peat) cease to be predicted at the larger sample spacings.

It was anticipated that the combination of expert knowledge with the quantitative sparse data mining would lead to higher map values compared to equivalent data mining models. This was found to be the case in the more complex Yeovil and Needwood Forest areas (Figure 58 and Figure 59) where the most valuable maps were created with the combined approach (see red circles on figures). Figure 64 to Figure 67 show the maps for Needwood Forest using the combined method incorporating data from the same samples used in Figure 60 to Figure 63. These maps show lower amounts of

misclassification than their pure data mining counterparts, and also more consistently predict the less extensive units which were lost with the data mining samples.

In the Worksop area, the addition of the expert knowledge to the sparse data mining did not improve on the map values achieved by just the data mining (Figure 57) except for the SOIL input beyond 1500 m sample spacing. The reason for this lack of improvement is that the geology and parent material distribution in the Worksop area is relatively simple, and appears to be better characterised by even the most sparse data mining sample (2800 m) than by the expert knowledge. This can be seen by comparing the level of the horizontal expert knowledge line (EK) with the data mining models (Figure 57).

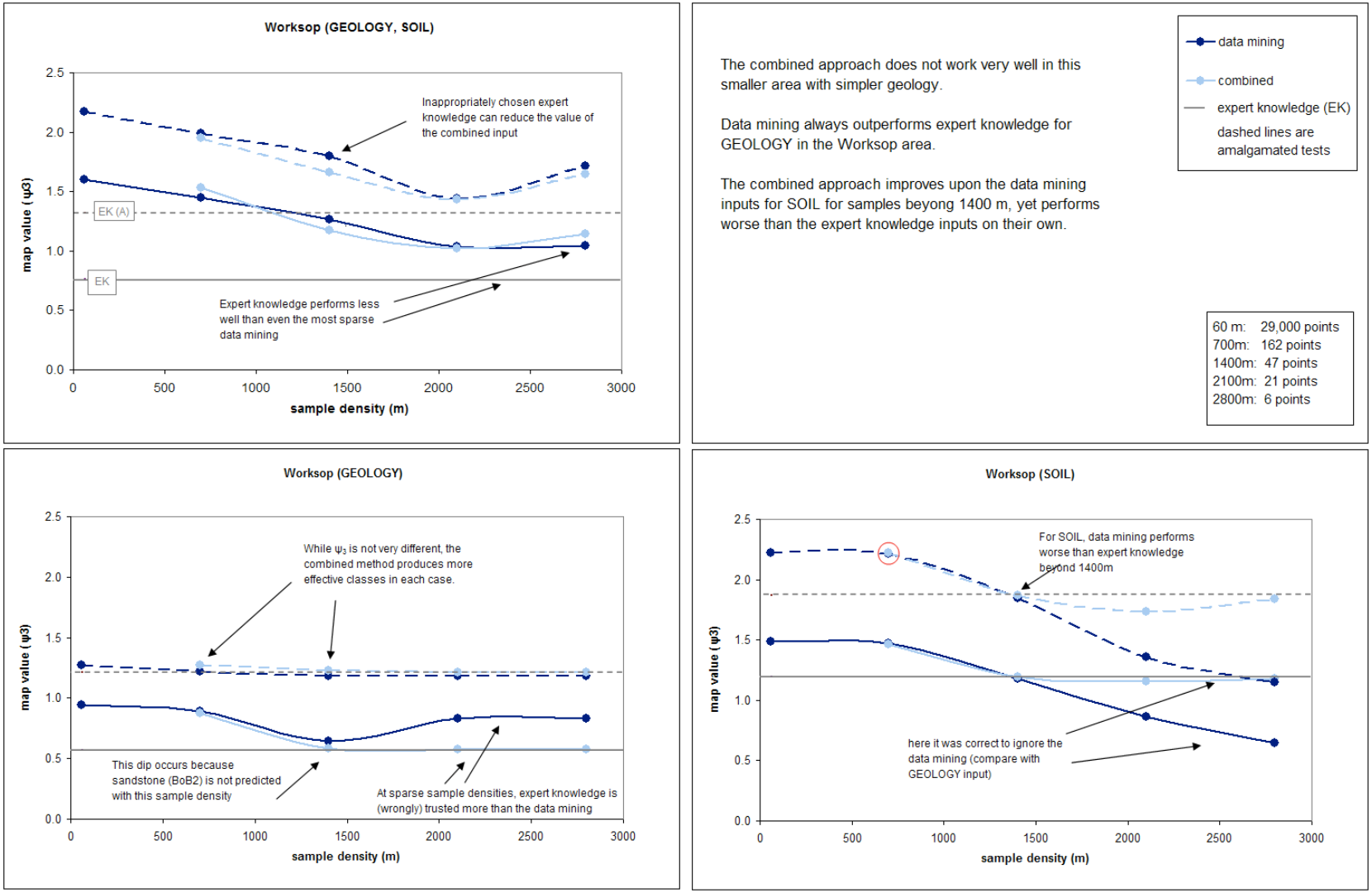


Figure 57 - Comparing the combined, data mining and expert knowledge approaches for the Worksop area

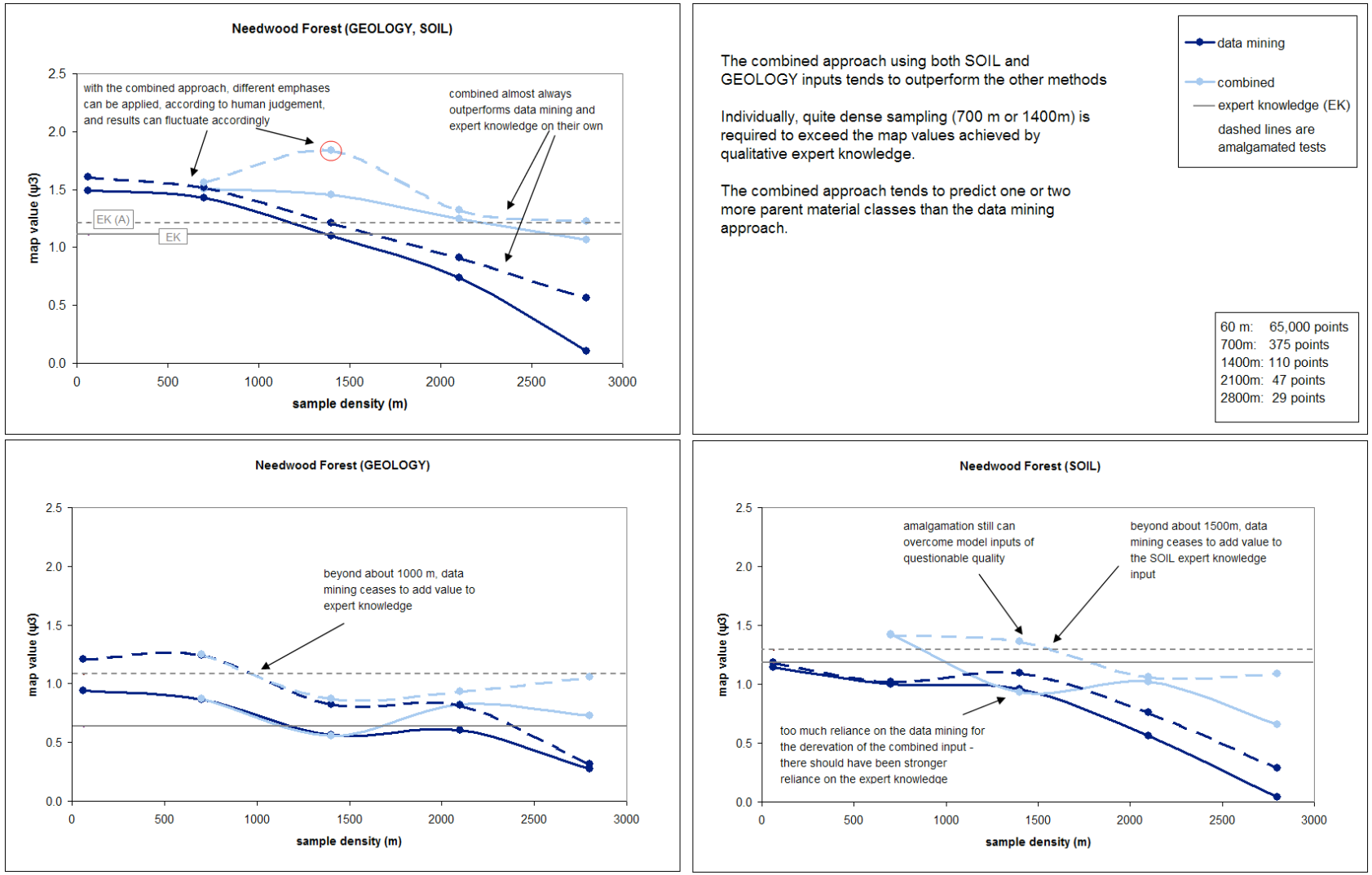


Figure 58 - Comparing the combined, data mining and expert knowledge approaches for the Needwood Forest area

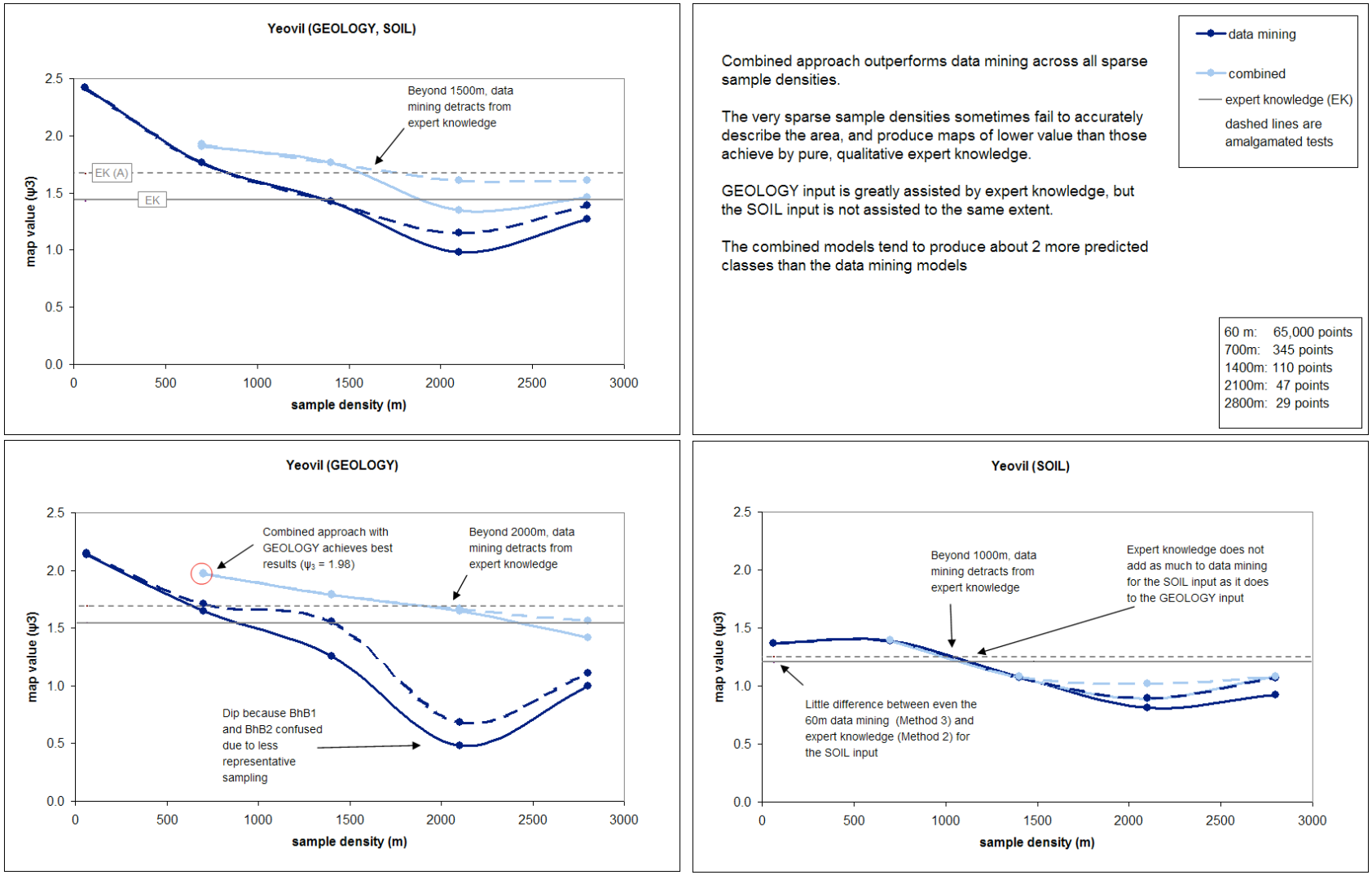


Figure 59 - Comparing the combined, data mining and expert knowledge approaches for the Yeovil area

Because the expert knowledge is comparatively poor in the Worksop area, the addition of expert knowledge to the sparse data mining had the effect of reducing the value of the resulting maps by a small amount, rather than increasing it.

In the Needwood Forest area, the combined approach using both GEOLOGY and SOIL inputs almost always outperformed the expert knowledge and data mining inputs on their own (Figure 58). However in the Yeovil area, beyond certain sample densities, the combined and data mining approaches do not perform as well as the qualitative expert knowledge derived inputs on their own (Figure 59). Indicative cut-off points (ranging between 1000 and 2000 m depending on the input) can be seen where the combined and data mining lines cross the horizontal grey expert knowledge lines. Similar patterns can be seen for the Worksop SOIL input, where a sample density above 1400 m is required to improve upon the expert knowledge input (Figure 57).

Across all study areas, a consistent effect of the combined approach was the increase in the number of effective classes (two additional classes is typical). This resulted from the additional prediction of smaller parent material classes, which typically were not sampled with the sparse data samplings. These extra classes are typically limited in extent and so do not greatly affect the achieved overall accuracies (θ_1). However, they do represent useful additions to the parent material models.

Because 1:25,000 scale reference soil parent material maps are more detailed than the 1:50,000 scale GEOLOGY and 1:250,000 scale SOIL predicting evidence layers, it was anticipated that smaller polygons seen on the reference map would be missing from the modelled outputs. This is demonstrated well in Figure 68, where the actual extent of each parent material class for the Worksop area is overlain (in translucent red) on the probability of prediction (in blue) from the model of Test W75. Frequently, the discrepancies arise in small polygons. Nevertheless the model output tends to predict the likely membership of the parent material classes reasonably well.

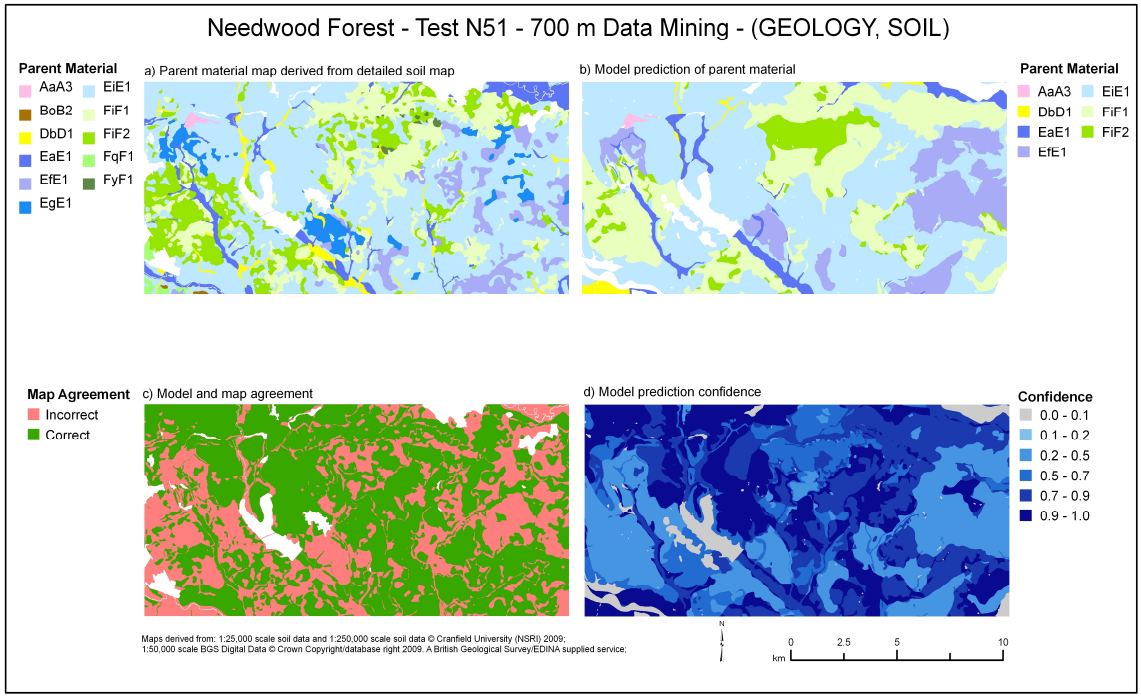


Figure 60 - Test N51 maps (700 m Data mining)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.43$; $\theta_1 = 0.62$; $C_e = 7$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

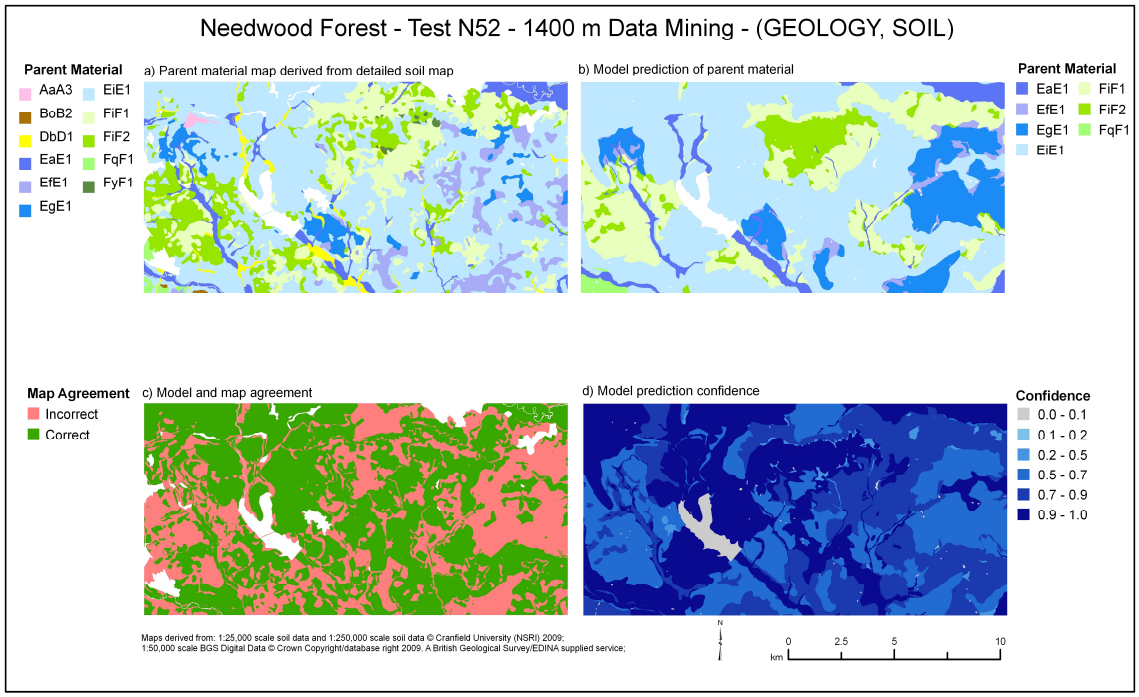


Figure 61 - Test N52 maps (1400 m Data mining)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.10$; $\theta_1 = 0.60$; $C_e = 7$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

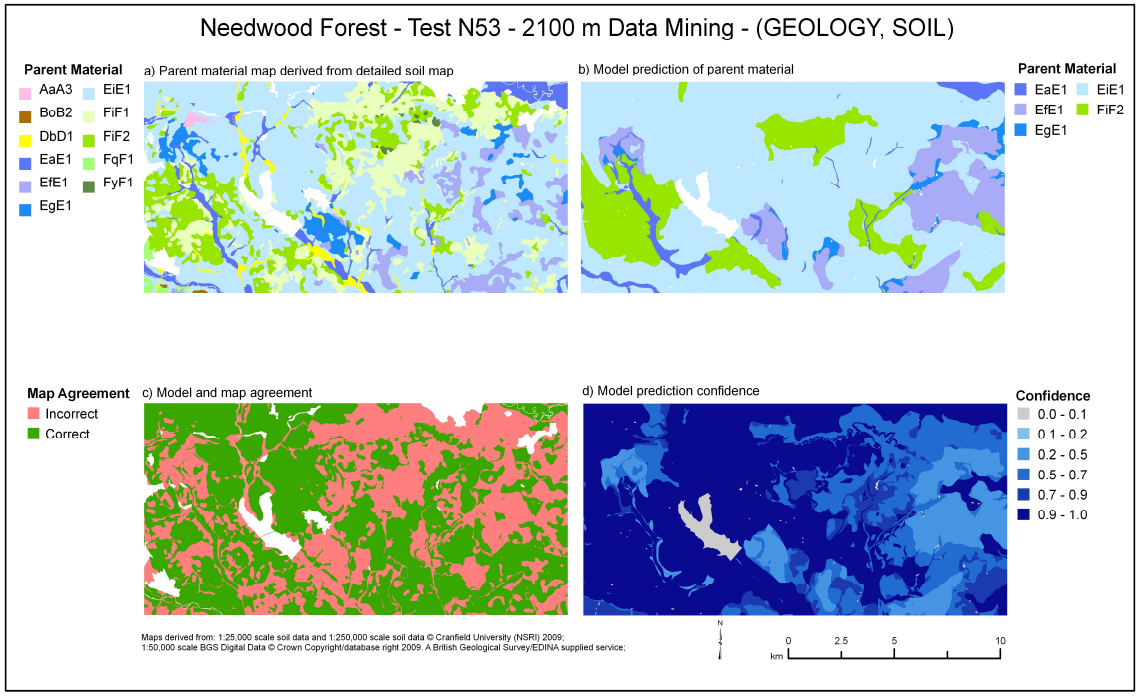


Figure 62 - Test N53 maps (2100 m Data mining)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 0.74$; $\theta_1 = 0.57$ $C_e = 5$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

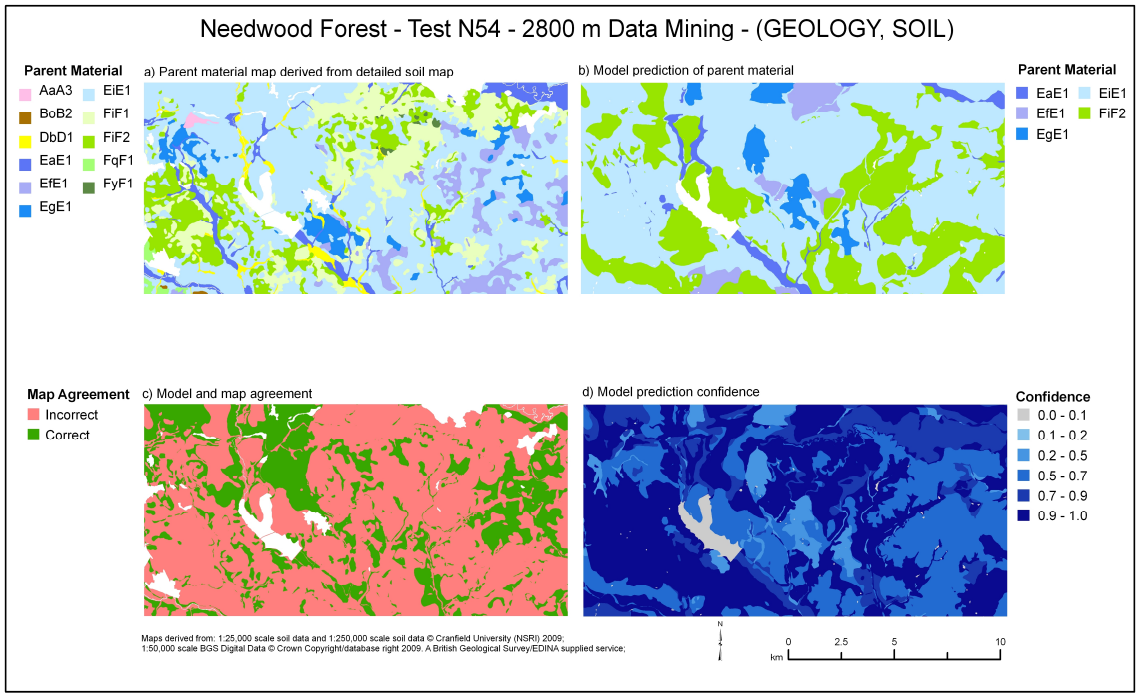


Figure 63 - Test N54 maps (2800 m Data mining)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 0.11$; $\theta_1 = 0.28$ $C_e = 5$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

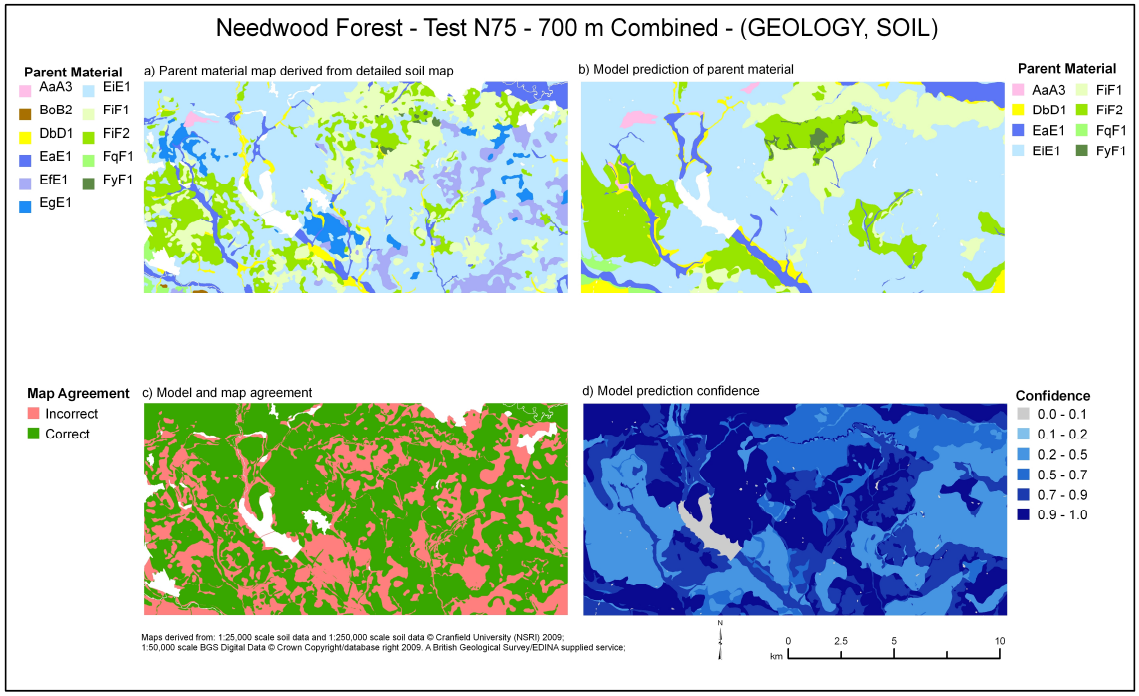


Figure 64 - Test N75 maps (700 m Combined)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.50$; $\theta_1 = 066$; $C_e = 8$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

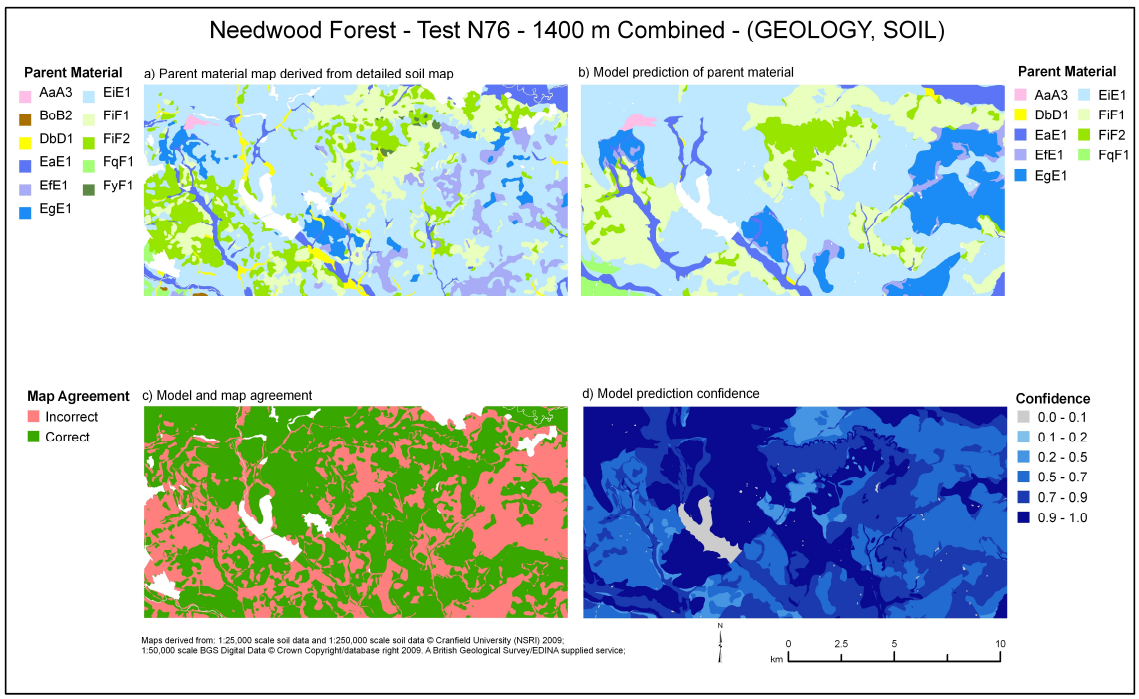


Figure 65 - Test N76maps (1400 m Combined)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.45$; $\theta_1 = 0.63$; $C_e = 9$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

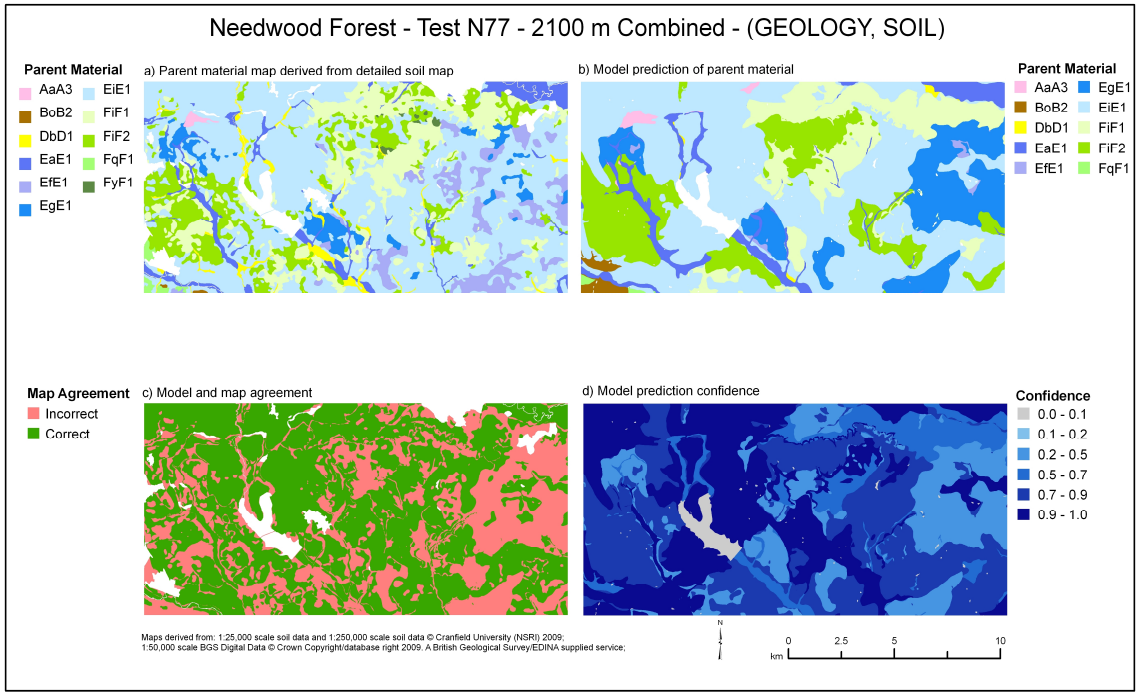


Figure 66 - Test N77 maps (2100 m Combined)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.24$; $\theta_1 = 0.62$; $C_e = 8$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

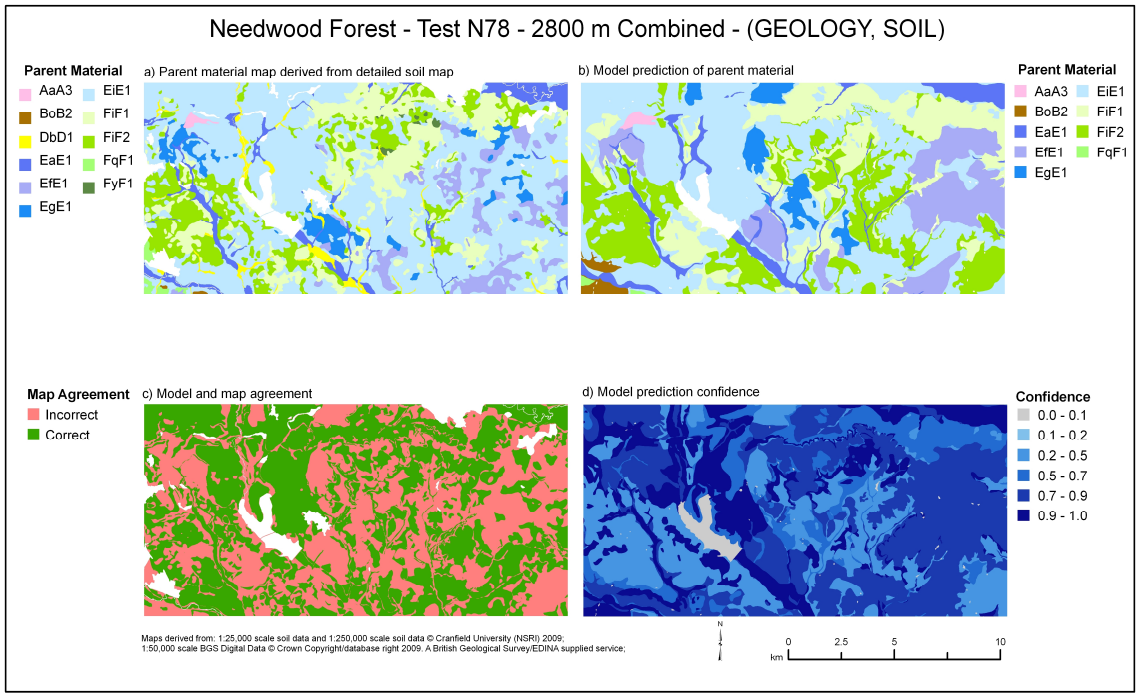


Figure 67 - Test N78maps (2800 m Combined)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.07$; $\theta_1 = 0.53$; $C_e = 9$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

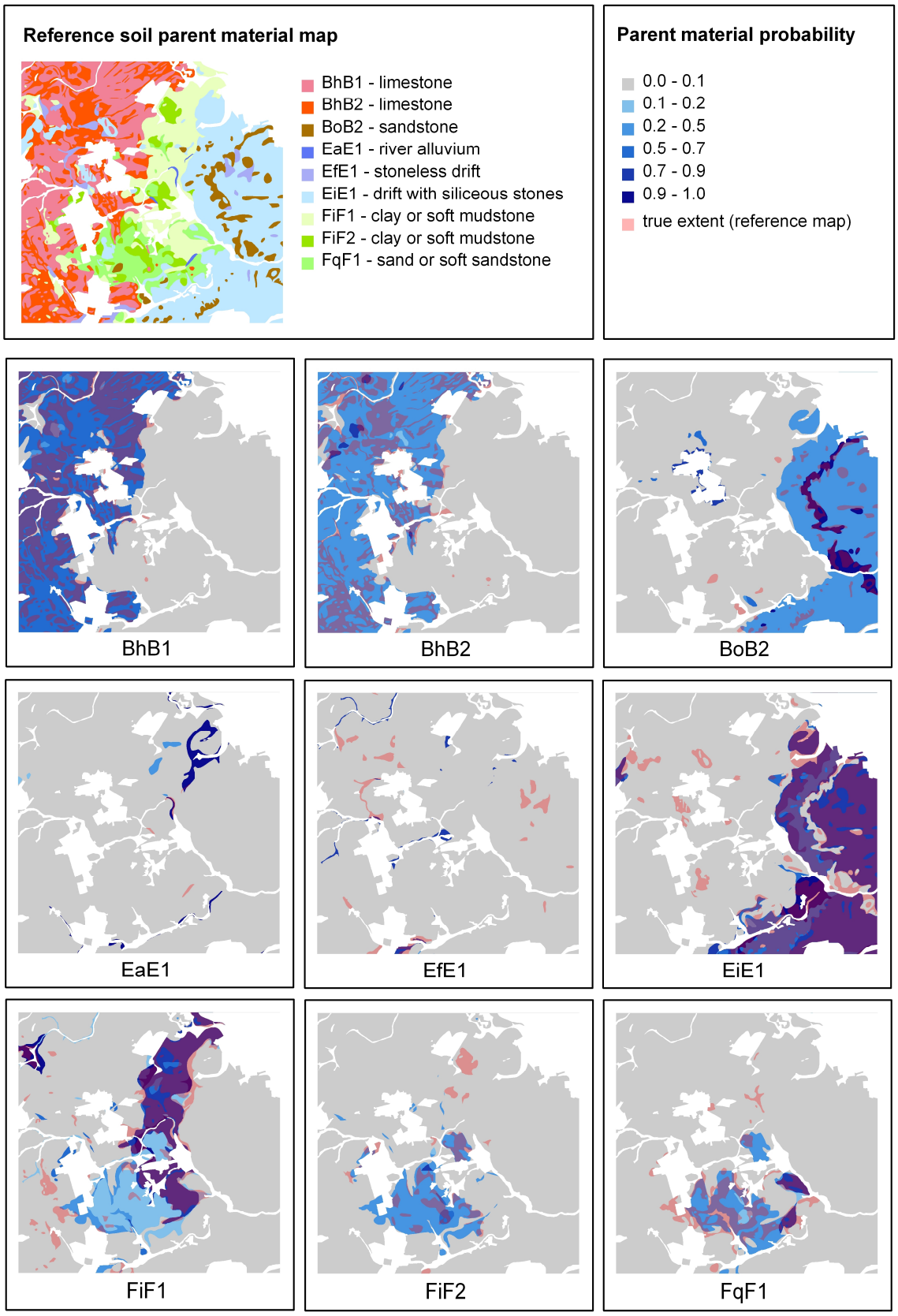


Figure 68 - Comparison of the predicted likely extent of each parent material class with the actual extent for Worksop, based on Test W75.
Note: The true extent is shown in translucent red, so over dark blue areas, this appears purple.

A visual comparison of the two limestone classes (BhB1 and BhB2; upper two maps in Figure 68; Test W75) clearly show a strong, if generalised, agreement between the model and the reference map. Nevertheless, because of their lithological similarities, consistent confusion remains. This confusion is unlikely to be overcome with the current evidence layers, and in such cases, class amalgamation is likely to be warranted. This can be seen by the improvement in map value shown in Figure 57, where the dashed (amalgamated) lines are noticeably higher than the unamalgamated lines. When only the SOIL input is considered, this amalgamation of the limestone units brings about an improvement in map value from 1.47 (W83, Figure 69) to 2.22 (W95, Figure 70) – the highest achieved in this study area. Additional evidence layers capable of describing the stoniness of the upper part of the soil profile are required to more accurately differentiate these limestone units.

8.6.1 Effective predictors of soil parent material

There was not a particular evidence layer which was the most successful in all study areas. The highest map value tended to be produced by the models using the SOIL input in the Worksop area, (Figure 57), by both the GEOLOGY and SOIL inputs in the Needwood Forest area (Figure 58) and by the GEOLOGY input on its own in the Yeovil area (Figure 59). Yeovil area is the only area in which the mapping of the National Soil Map predates the detailed mapping of the soil from which the reference soil parent material map was created.

Future work should consider the wider application of this finding; specifically considering other areas, which were unmapped in detail prior to the creation of the National Soil Map (the SOIL layer). Further tests should investigate whether SOIL adds significant detail or value to the maps of soil parent material which can be created based solely upon the geological mapping.

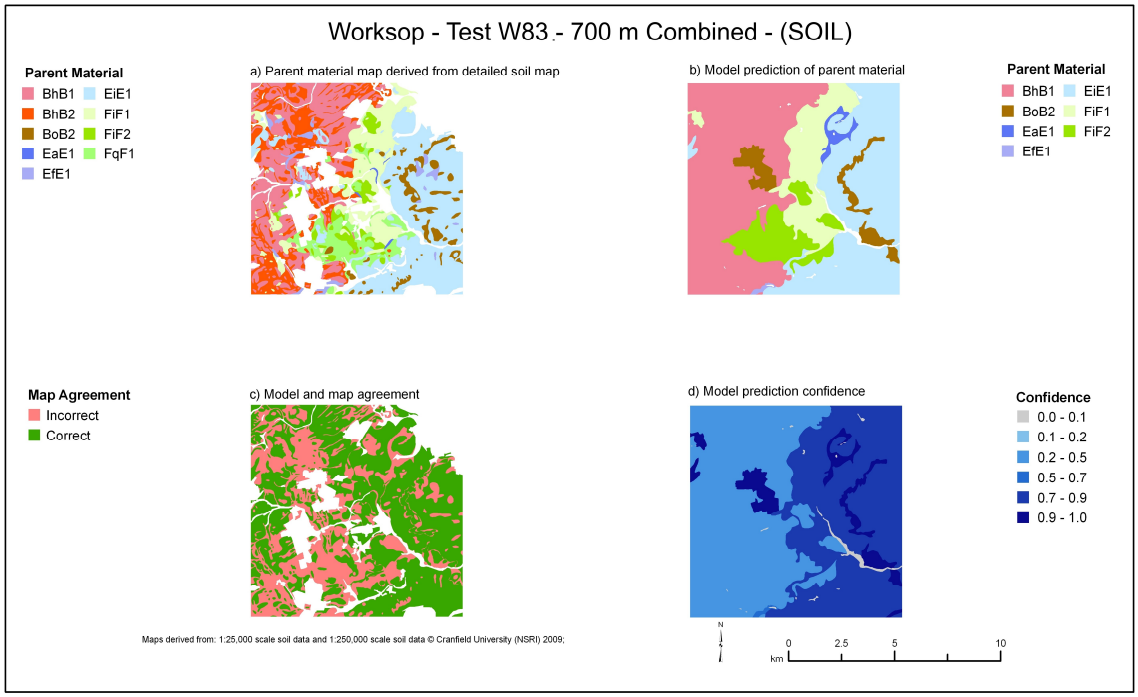


Figure 69 - Test W83 maps (Combined methodology – 700 m sample)
Inputs: SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.47$; $\theta_1 = 0.63$ $C_e = 7$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

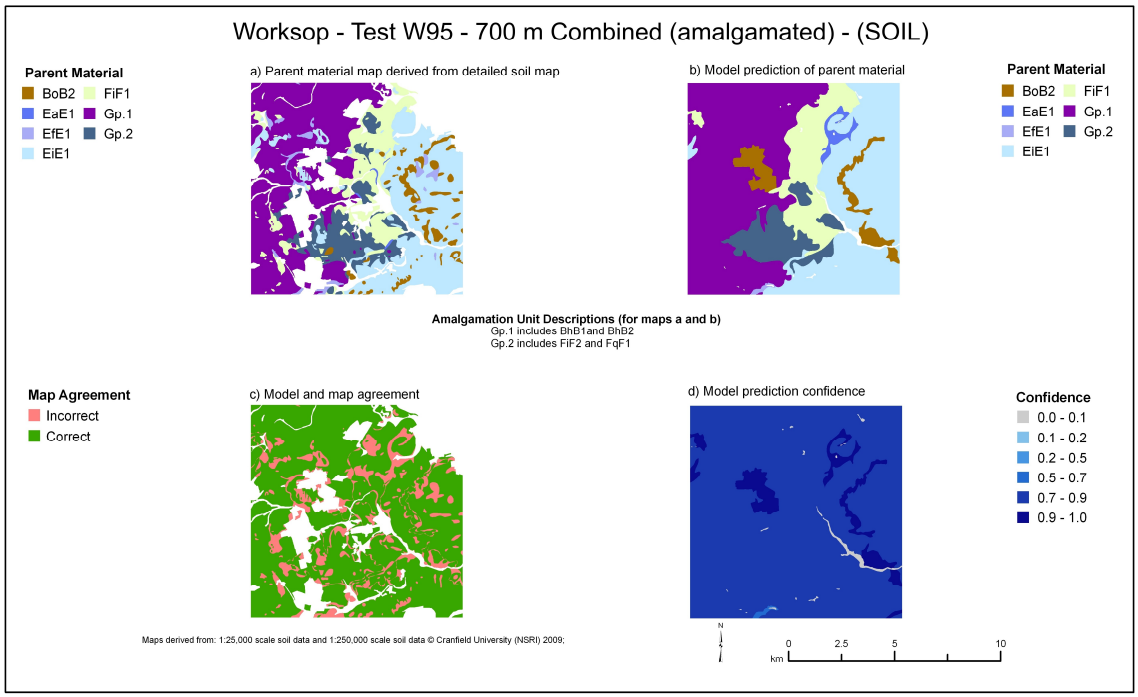


Figure 70 - Test W95 maps (Combined methodology – 700 m sample)
Inputs: SOIL; Classification: Amalgamated NSRI PARLITH; $\Psi_3 = 2.22$; $\theta_1 = 0.84$ $C_e = 7$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

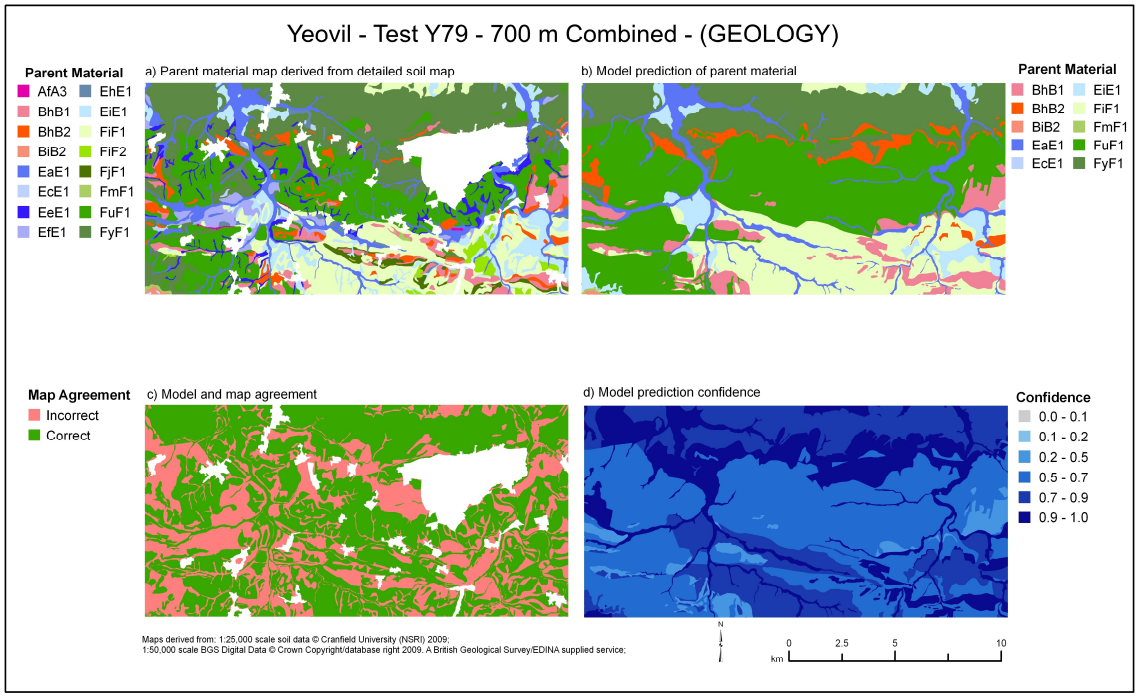


Figure 71 - Test Y79 maps (Combined methodology – 700 m sample)
Inputs: GEOLOGY; Classification: NSRI PARLITH; $\Psi_3 = 1.97$; $\theta_1 = 0.63$ $C_e = 9$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

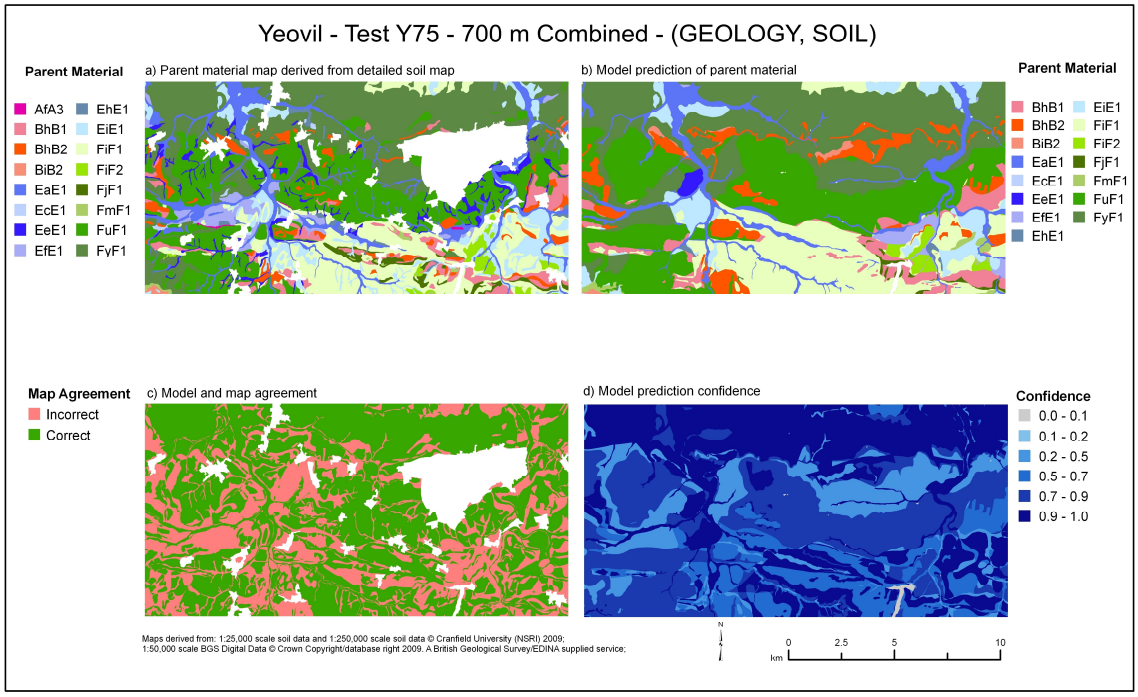


Figure 72 - Test Y75 maps (Combined methodology – 700 m sample)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.91$; $\theta_1 = 0.65$ $C_e = 14$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

The tests with more effective classes (C_e) tend to be those using both GEOLOGY and SOIL evidence layers. This is because when the model is only using one input, the same class will be predicted every time the same predicting class is present. With a model with multiple inputs, the intersection of multiple probabilities can lead to the prediction of additional parent materials based on the combination of predicting classes present. This holds true even in areas where the highest map values were achieved with single inputs. For example, the 700 m combined GEOLOGY input achieved a map value of 1.97 with 9 effective classes (Y79, Figure 71). The equivalent test using both GEOLOGY and SOIL (Y75, Figure 72) achieved a slightly lower map value of 1.91, but had 14 effective classes, and a slightly higher overall accuracy (0.65 vs. 0.63).

8.6.2 Map success and the complexity of class membership

Any judgement of the success of the predicted maps will always depend on the purpose to which the results will be put. The examination and assessment of the model outputs in this research has focussed on the creation of a traditional class based map product, with the aim of accurately predicting the most likely parent material at any given point. However, consistently accurate prediction of individual parent material classes, based on the available environmental correlatives, has been shown to be difficult. High overall accuracies (θ_1) are achievable where broad amalgamated classes are used, for example, test N91, (θ_1= 0.90, C_e = 4) where 4 parent material units are amalgamated into one parent material map unit. While N91 achieves a high overall accuracy, it does not achieve the highest map value (ψ_3) for maps using the 700 m sample. This is achieved by creating two amalgamated classes with 3 and 2 members respectively (N87, Figure 74, θ_1= 0.83, C_e = 7). Comparing the map of agreement (Figure 74c) with that of the equivalent unamalgamated test (N75, Figure 73c), the improvement (reduction in extent of red) is noticeable.

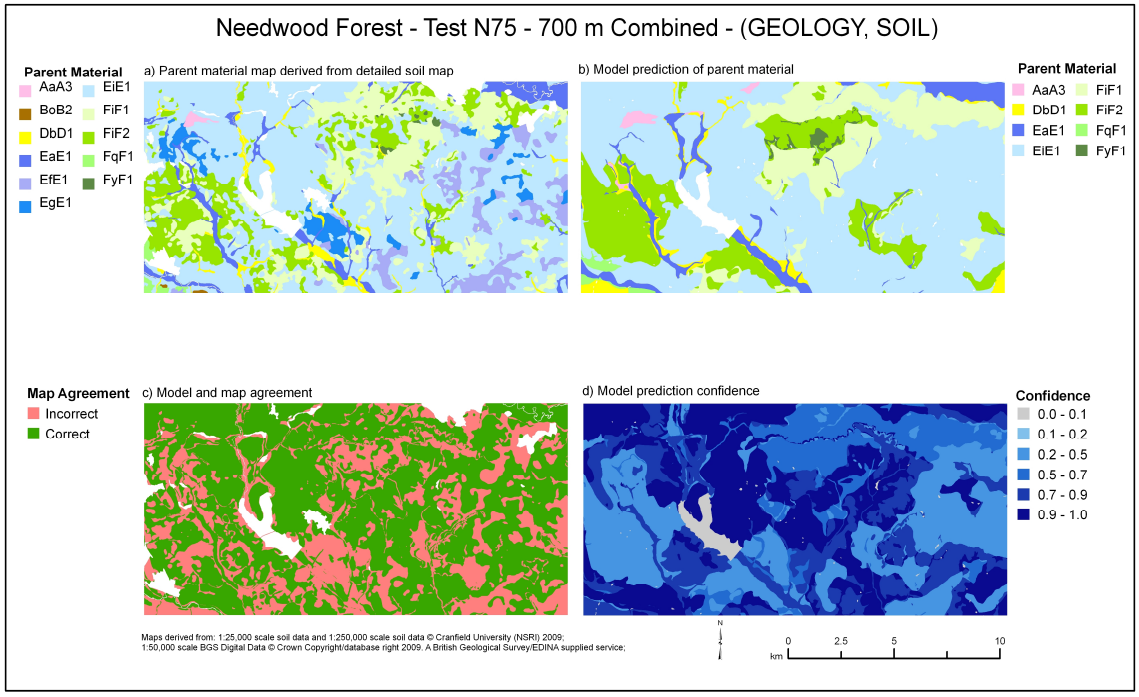


Figure 73 - Test N75 maps (Combined methodology – 700 m sample)
Inputs: GEOLOGY, SOIL; Classification: NSRI PARLITH; $\Psi_3 = 1.50$; $\theta_1 = 0.66$ $C_e = 8$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

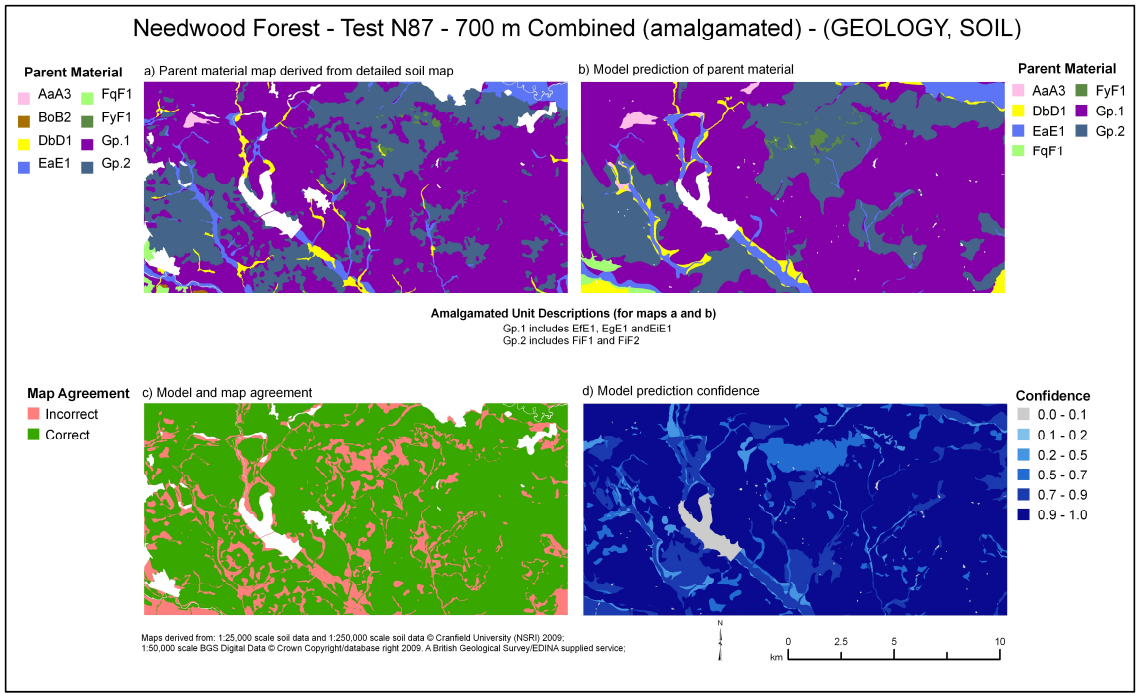


Figure 74 - Test N87 maps (Combined methodology – 700 m sample)
Inputs: GEOLOGY, SOIL; Classification: Amalgamated NSRI PARLITH; $\Psi_3 = 1.54$; $\theta_1 = 0.83$ $C_e = 7$
A larger version is available in Appendix 4. NSRI codes are described in Appendix 2.

The matrices comparing parent material with the predicting classes from the expert knowledge, data mining and combined approaches all show a degree of fuzziness in the relationships. Rarely is there a 1:1 relationship between a predicting evidence layer and a soil parent material class. Thus, to make full use of the resulting maps from this type of model, users should be made aware of the uncertainty of the predicted parent material classes, and, where possible, take account of any sub-dominant classes predicted by the model. For this reason, even if class amalgamation is used, it is beneficial to include the full detail of the probabilities of each parent material unit, given the predicting evidence layers.

8.6.3 Consistently successful parent material classes

The class value (ξ) of the different parent materials units obtained using the expert knowledge, data mining and the 700 m combined methodologies are compared in Table 49 to Table 47. The class value assessment provides an indicator of the overall usefulness of each parent material class, based on the amount of over and under prediction, as well as the proportion of accurate prediction (see section 4.5.1.2, p 75). These tables have been sorted from the highest to lowest class value of the 700 m combined methodology.

In the Worksop area, (Table 49) only the combined methodology predicts all parent materials. In Needwood Forest and Yeovil, some minor units were never predicted.

In all study areas, it is the widespread parent material classes that are predicted with the highest class values (Table 49 to Table 47). These relationships for the three study areas have been plotted in Figure 75 to Figure 78 using results from the expert knowledge, data mining and 700 m combined methods. To aid discussion, indicative trend lines have been fitted through the points from the data mining methodology. This allows identification of units which do not follow this general relationship.

Table 49 - The prediction of parent material classes from the expert knowledge, data mining and 700 m combined methods for Worksop

		Expert Knowledge Methodology			Data Mining Methodology			700 m Combined Methodology		
P.M.	**% area**	**ξ max**	**evidence**	**confusion**	**ξ max**	**evidence**	**confusion**	**ξ max**	**evidence**	**confusion**
EiE1	30%	0.87	SOIL	BoB2	0.87	SOIL	BoB2	0.87	SOIL	BoB2
FiF1	13%	0.76	SOIL	FqF1	0.77	SOIL, SLOPE	BhB1, FqF1	0.77	SOIL	FiF2, FqF1, BhB1
BhB1	22%	0.72	SOIL	BhB2	0.72	SOIL	BhB2	0.72	SOIL	BhB2
FqF1	7%			FiF1, BoB2	0.48	SOIL, SLOPE	FiF2	0.59	SOIL, GEOLOGY	FiF2, FqF1, BhB1
BoB2	5%	0.28	SOIL	FqF1, EiE1	0.53	SOIL, GEOLOGY	EiE1	0.50	SOIL, GEOLOGY	EiE1
FiF2	4%			FiF1, BoB2	0.40	All	FiF1, FqF1	0.38	SOIL	FqF1
EfE1	2%	0.03	GEOLOGY, SLOPE	BhB1, EiE1	0.24	SOIL	BhB1, EiE1	0.24	SOIL	FiF1, EiE1, BhB1
EaE1	0.20%	0.17	All	FiF1			FiF1	0.11	GEOLOGY	FiF1, EiE1
BhB2	15%	0.54	GEOLOGY, SLOPE	BhB1	0.15	SOIL, GEOLOGY	BhB1	0.10	SOIL, GEOLOGY	BhB1

Notes for Table 49,Table 50 and Table 47 **P.M.** -parent material**. % areas** – % of study area covered in that unit, according to the reference maps. **ξ max** - the highest class value achieved using any evidence layers. **evidence** – the evidence layers which resulted in the highest class value. **confusion** - units commonly confused with the parent material listed in the P.M. column.

Table 50 - The prediction of parent material classes from the expert knowledge, data mining and 700 m combined methods for Needwood Forest

		Expert Knowledge Methodology			Data Mining Methodology			700 m Combined Methodology		
P.M.	% area	ξ max	evidence	confusion	ξ max	evidence	confusion	ξ max	evidence	confusion
EiE1	57%	0.76	SOIL	EfE1, FiF1,FiF2	0.78	SOIL	Most units	0.80	SOIL, GEOLOGY	Most units
AaA3	0.20%	0.71	GEOLOGY	EiE1	0.73	SOIL, GEOLOGY	EiE1	0.70	GEOLOGY	EiE1
EaE1	8%	0.53	SOIL, GEOLOGY	EiE1, DbD1	0.60	SOIL, GEOLOGY	EiE1, DbD1	0.67	SOIL, GEOLOGY	DbD1, EiE1
FiF1	12%	0.53	SOIL	EiE1,FiF2	0.53	SOIL, GEOLOGY	EiE1,FiF2	0.64	SOIL, GEOLOGY	EiE1,FiF2
FiF2	10%	0.45	SOIL	EiE1, FiF2	0.45	SOIL	EiE1, FiF2	0.55	SOIL	EiE1, FiF1
FqF1	0.03%	0.18	GEOLOGY, SLOPE	EaE, EiE1	0.37	SOIL, GEOLOGY	EiE1	0.23	SOIL, GEOLOGY	EiE1
FyF1	0.40%			EiE1, FiF1, FiF2	0.01	All	EiE1, FiF1, FiF2	0.13	SOIL, GEOLOGY	FiF2
DbD1	3%			EaE1, EiE1			EaE1, EiE1	0.09	SOIL, GEOLOGY	EiE1, FqF1
BoB2	0.01%			EaE1, EiE1	0.38	SOIL, GEOLOGY	EiE1			DbD1
EfE1	7%	0.47	SOIL	EgE1, EiE1	0.46	SOIL, GEOLOGY	EgE1, EiE1			EiE1
EgE1	4%			EfE1			EfE1, EiE1			EiE1

Table 51 - The prediction of parent material classes from the expert knowledge, data mining and 700 m combined methods for Yeovil

		Expert Knowledge Methodology			Data Mining Methodology			700 m Combined Methodology		
P.M.		ξ max	evidence	confusion	ξ max	evidence	confusion	ξ max	evidence	confusion
FyF1	29%	0.75	SOIL	FuF1, EaE1	0.78	SOIL, GEOLOGY	FuF1, EaE1, EfE1, EiE1	0.77	SOIL	FuF1, EaE1, EfE1
FuF1	23%	0.72	SOIL	FyF1, EeE1	0.75	SOIL, GEOLOGY	FyF1, EeE1 + others	0.75	SOIL	EeE1, FyF1 (+ others)
FiF1	14%	0.66	GEOLOGY, SLOPE	EiE1 (+ others)	0.7	GEOLOGY	EiE1 (and others)	0.71	SOIL	EiE1 (and most others)
EaE1	9%	0.69	GEOLOGY, SLOPE	FyF1, EiE1, FiF1	0.69	GEOLOGY	FyF1, EiE1, FiF1	0.69	GEOLOGY	FuF1, FyF1
BiB2	0%	0.76	SOIL, GEOLOGY	EiE1	0.80	GEOLOGY, SLOPE	EiE1, FmF1	0.65	GEOLOGY	FmF1, FyF1
BhB1	5%	0.49	GEOLOGY	FiF1, BhB2	0.56	SOIL, GEOLOGY	BhB2, FiF1, FuF1	0.54	GEOLOGY	BhB2
EiE1	7%	0.31	SOIL	EhE1, FiF1,FyF1	0.50	All	FiF1, FyF1	0.47	SOIL	FiF1, FyF1
FiF2	1%			FiF1	0.43	SOIL, GEOLOGY	EiE1,FiF1,FiF2	0.42	SOIL	FiF1
BhB2	4%	0.31	SOIL, GEOLOGY	BhB1, FyF1	0.32	SOIL, GEOLOGY	BhB1, FuF1, FiF1	0.37	SOIL	BhB1, FuF1
FmF1	0%	0.17	GEOLOGY	BhB1	0.26	GEOLOGY	BiB2	0.18	GEOLOGY	EiE1, FiF1
FjF1	1%	0.03	All	FiF1	0.31	All	FiF1, BhB1	0.18	SOIL	FiF1
EfE1	2%	0.07	GEOLOGY, SLOPE	EiE1, FyF1, FuF1	0.06	All	FyF1, FiF1, EiE1	0.08	SOIL	EaE1, EiE1, FyF1
EeE1	4%	0.01	GEOLOGY	EiE1, FyF1, FuF1			FuF1, FyF1, FiF1	0.05	SOIL, GEOLOGY	FuF1, FyF1
AfA3	0%	0.00		FuF1			FuF1, EaE1			FuF1
EcE1	0%			FuF1, FyF1, FiF1			FuF1, FyF1			FuF1, FyF1
EhE1	1%	0.05	SOIL	EiE1, FyF1	0.12	SOIL, GEOLOGY	FiF1, FyF1			FiF1, FyF1

8.6.3.1 Parent material success in the Worksop area

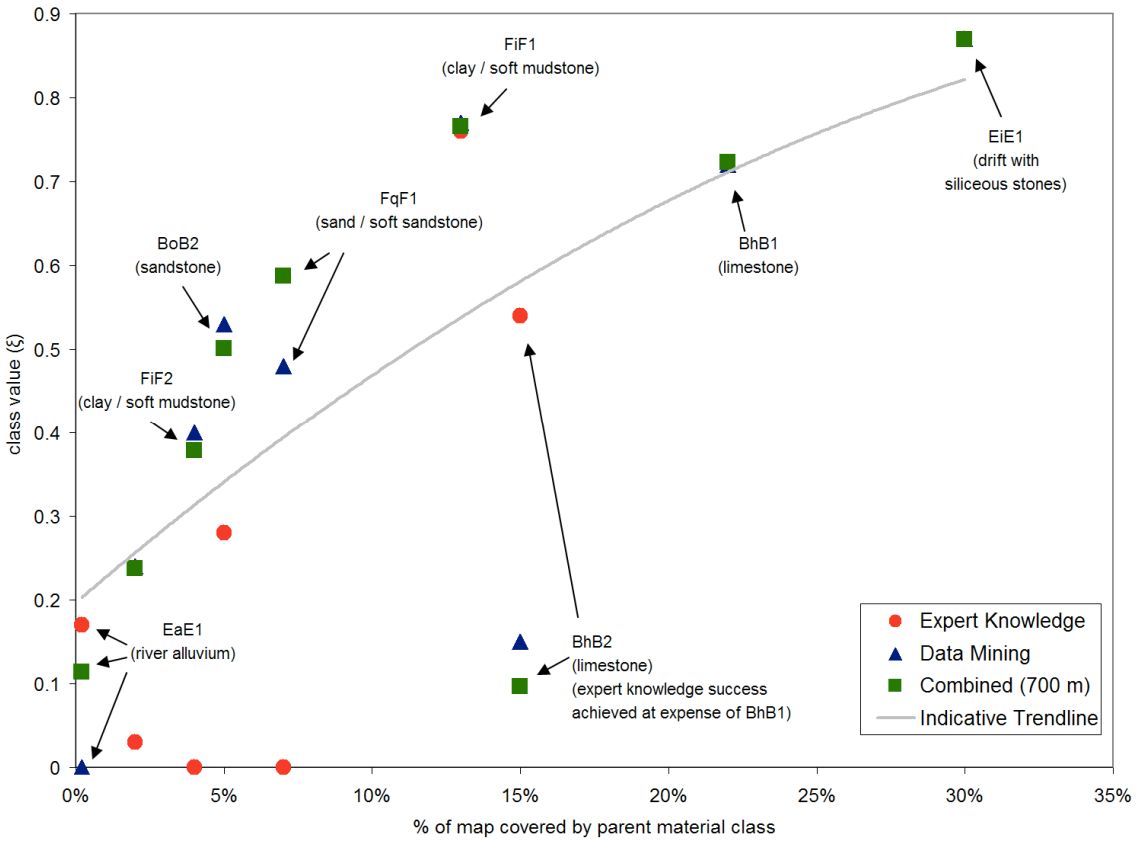


Figure 75 - Parent material success plotted against area of map (Worksop)
Note: The best prediction from the evidence layer combinations for each method is plotted. The BhB1 and BhB2 layers predicted using the expert knowledge method were achieved using different predictors, and achieved at each other's expense.

All parent materials were predicted using the 700 m combined methodology in the Worksop area. Although three of the nine classes had ξ values less that 0.25, 65% of the area of the map was well predicted by parent material classes with ξ of over 0.70. The most extensive units in the Worksop area are also those which achieve the highest class values. The most notable outlier in the Worksop area is the poor performance of BhB2 (limestone – soil over lithoskeletal substrate) which achieves, at best, a class value (ξ) of 0.10. This parent material unit occupies approximately 15% of the area, making it the third most extensive class, according to the reference map. This unit is often confused with BhB1 (limestone – soils with lithoskeletal substrate). There is a subtle difference between these parent materials. The descriptions of these materials, from (Clayden and Hollis, 1984) are compared below.

Lithoskeletal soils (B1) are those in which bedrock or angular skeletal material occupies at least half of the upper 80 cm of the profile. In addition, they have no surface layer more than 30 cm thick that contains less than 16% stones by volume.

Soils over lithoskeletal material (B2) have one of the following types of surface layer:
(a) at least 30 cm of material containing less than 16% stones by volume
(b) at least 40 cm of material containing less than 36% stones by volume.

This type of detail tends not to be recorded by geological mapping as these types of differentiating characteristics have a stronger emphasis on the soil side of the soil-parent material-geology continuum.

FiF1 (clay or soft mudstone) performs slightly better than might be anticipated by its extent. This is due, predominantly to the extent of the BROCKHURST 2 (711c) map unit of the National Soil Map (SOIL) evidence layer. As the National Map was created after the detailed soil map, it is unknown how useful this class would be at predicting parent material in a region previously unmapped in detail. Nevertheless, clay and soft mudstone is a parent material which may be readily identified by both geologists and soil surveyors.

8.6.3.2 Parent material success in the Needwood Forest area

The Needwood Forest area is dominated by the drift with siliceous stones parent material, which was predicted quite well (EiE1, ξ = 0.80, Figure 76 and Table 50). Other well-predicted units include peat (AaA3, ξ = 0.70) and river alluvium (EaE1, ξ = 0.67). These last two parent materials are distinctive in composition and occupy easily identified low ground in the landscape, allowing reasonably consistent mapping between the geological and soil surveys. For smaller units like the peat, the 1:50,000 scale GEOLOGY layer outperforms the 1:250,000 SOIL layer, as it more accurately and consistently delineates the peat at a more detailed scale.

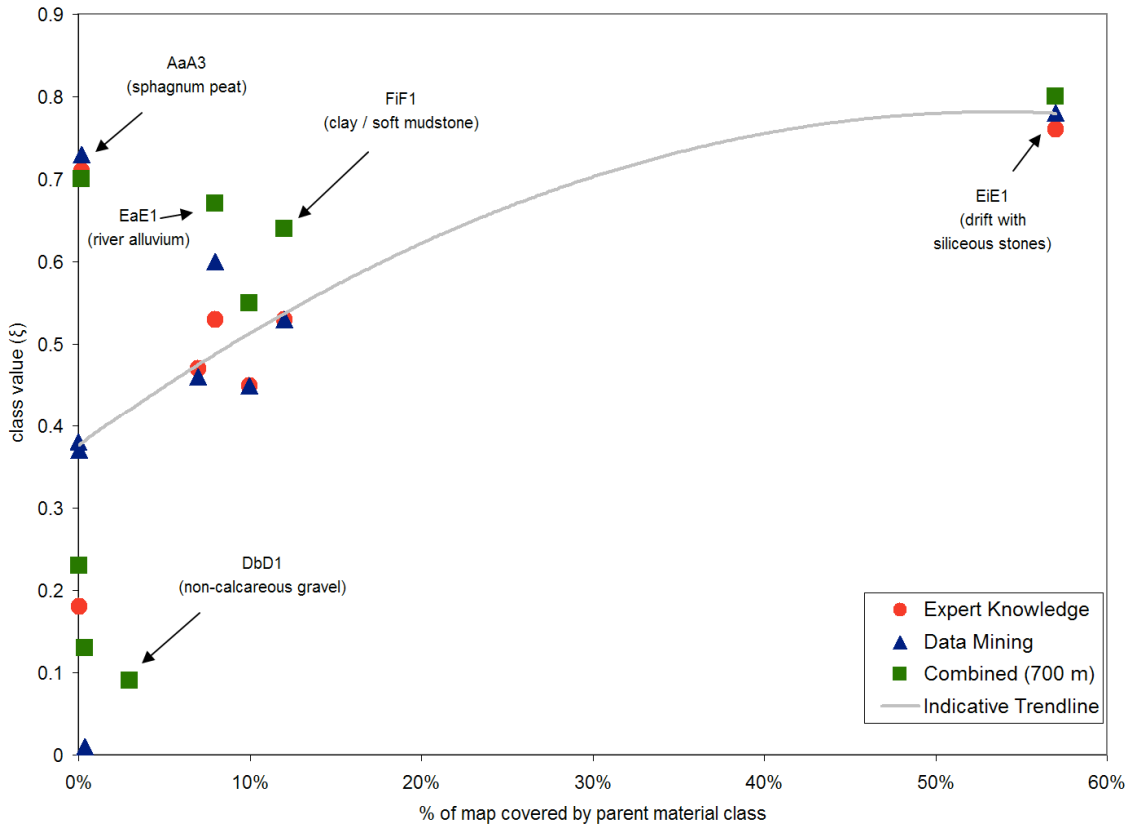


Figure 76 - Parent material success plotted against area of map (Needwood Forest)
Note: The best prediction from the evidence layer combinations for each method is plotted.

Again, the easily recognisable clay / mudstone unit (FiF1) performs reasonably well ($\xi = 0.64$) in this study area. However, some units do not perform well. Non-calcareous gravel (DbD1) performs a little worse than might be hoped, but, given its coverage of only 3% of the area such units may be overlooked by the less detailed evidence layers. More importantly, three parent materials which, together, amount to more than 10% of the area are not predicted at all. EfE1 (stoneless thick drift) and EgE1 (chalky thick drift) are not predicted by the combined method and yet previously EfE1 had been moderately well predicted by the Expert Knowledge method ($\xi = 0.47$). Many of these discrepancies depend on the amount of trust given to the qualitative and quantitative information to the combined inputs, and highlight the point that in this system involving expert judgement, better or worse choices can be made.

Drift deposits were shown to be difficult to differentiate, particularly when they occur in small units not reflected by the geological map. For example, compare the rather

fragmented distribution of EfE1 on the reference map (Figure 11, p51) with the broader extent of till (TILMP-DMTN) shown on the GEOLOGY or SOIL layers (Appendix 1) where EiE1 (thick drift with siliceous stones) is predicted.

Figure 77 provides a simple comparison of the extent of drift shown by the geological map (blue) compared with the thick drift (greater than 80 cm thickness, shown in green) and all drift (beige) according to the reference parent material map. It can clearly be seen that there is greater association between the thick drift and the geological mapping than the thin drift. Nevertheless, the geological mapping still only accounts for approximately 83% of the thick drift, according to the reference map. This could be due to the scale differences of the mapping (1:50,000 GEOLOGY vs. 1:25,000 reference map) and also to do with the differing mapping priorities of the geological and soil survey.

Where the evidence layers do not contain enough detail to pick out the smaller parent material units, and no additional datasets are available to aid prediction, the creation of amalgamated classes may be the optimum strategy. In this case however, as EiE1 is so dominant, the highest map values were not achieved when the classes were amalgamated. Thus, in cases such as these, defining the parent material map units by a name may be of use, but fundamentally, it is of more use to know that within the EiE1 map unit, 70% of the are is likely to be EiE1, while 20% is likely to be EfE1 and 10% EgE1. Such proportions might be estimated from the model probabilities or by analysis of the confusion matrices. This understanding of the fuzziness of the classification can then be fed forward into any further applications.

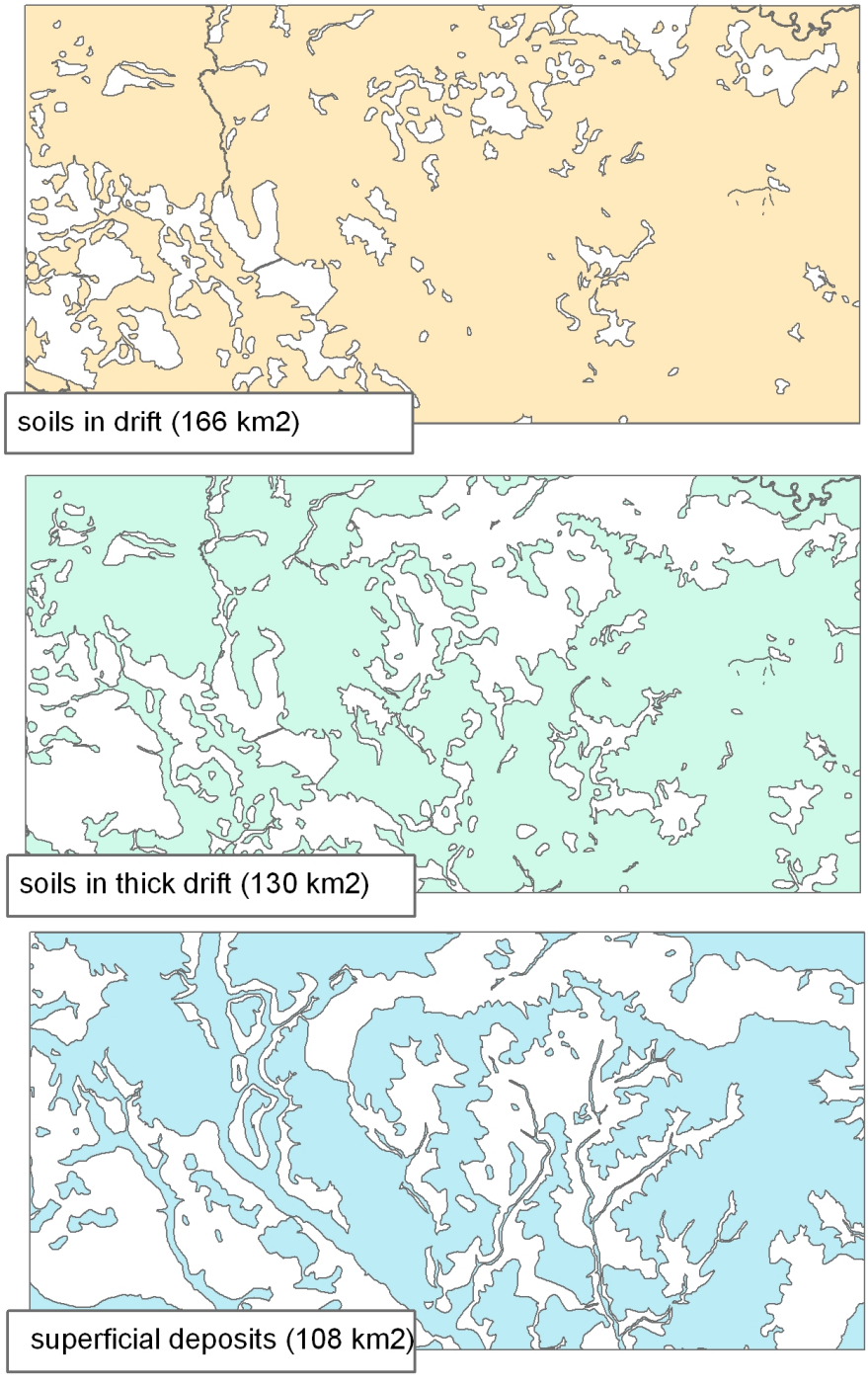


Figure 77 - Comparing drift extents in the Needwood Forest area
Note: the top two maps are derived from the reference soil parent material map. The bottom map is derived from the superficial geology layer (GEOLOGY). The geology layer under represents the extent of the superficial deposits relative to the reference map.

8.6.3.3 Parent material success in the Yeovil area

There is a good agreement in the Yeovil area between the extent of the parent material unit on the reference soil map and the class value (Figure 78). Units which perform notably better than expected include chalk (BiB2) and river alluvium (EaE1). Once again, these are physically or chemically distinctive units.

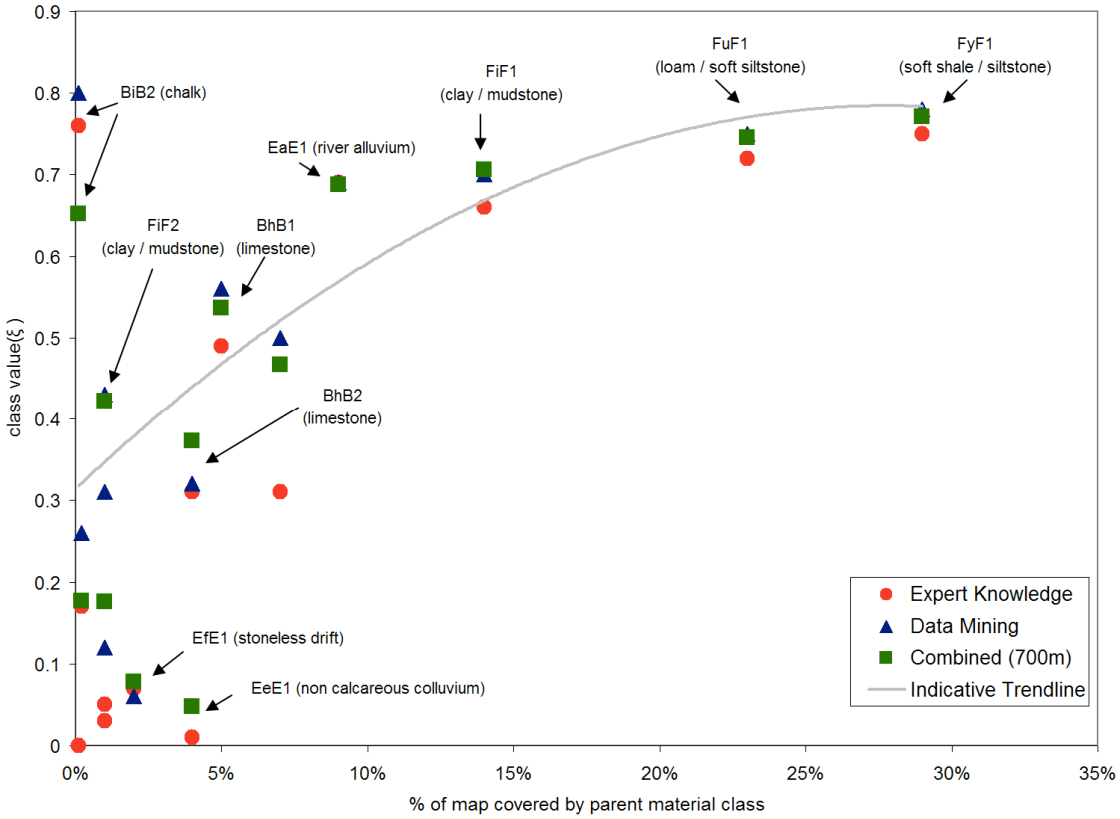


Figure 78 – Parent material success plotted against area of map (Yeovil)
Note: The best prediction from the evidence layer combinations for each method is plotted.

Stoneless drift (EfE1) and non calcareous colluvium (EeE1) perform less well than might be expected, although making up less than 5% of the area each, these units are not extensive and might easily, for the sake of cartographic clarity, be overlooked by those mapping at 1:50,000 and 1:250,000 scale. However, comparing these classes with limestone (BhB1, BhB2) which occupy similar extents on the reference map, there is a noticeable difference in the achieved class value. Limestone is more easily identified by geologists than such drift deposits.

8.6.4 Parent material classes and proportions

A clear trend in all areas has been demonstrated between extensive units and well-predicted units. Nevertheless, some parent materials achieve higher class values than might be expected. These classes are typically chemically, physically or morphologically distinctive, such as peat or chalk or have an expression in the landscape which is easy to map, such as alluvium.

8.6.5 Method transferability and scalability

While the issue of extrapolation lies beyond the scope of this reseach project, it is benefitial to consider some of the issues which are likely to empact upon the transferability and scalability of these methods when applied to new areas. These include the availability of datasets, the extent of similar landscapes, the geodiversity of the landscapes, and issues of scale.

The datasets (SLOPE, SOIL, GEOLOGY) which have been used in this research are available across all of the UK. Each of the map units for these datasets can be extended to their maximum extent from the three study areas, thus broadly defining the extent of similar landscapes to the study areas. As the locations of the detailed soil maps were chosen to be representative of the surrounding soil-landscapes, it is likely that such landscapes will extend for some distance beyond the study area. Initial tests have indicated that the model inputs developed on the test areas will broadly apply to the surrounding areas, with a fair degree of classification success.

These techniques are likely to work on other areas of England and Wales proximal to available reference detailed soil mapping (Figure 3, page 3). While upland areas were not assessed in this research, it is likely that these techniques will perform as well, if not better in such landscapes. In upland settings there are stronger controls by the landscape, so it is likely the slope or other landscape attributes will be a more effective predictor of parent material in such landscapes. For example, drift deposits are more

likely to be found in valleys than on steep slopes. Furthermore, due to te greater geo-diversity found in such regions, there will be greater taxonomic distances between the parent material types present in such upland areas, particularly between such diverse materials as peat and igneous intrusions. In contrast, the parent materials investigated in this research have been quite similar due their derivation from sedimentary and quaternary deposits.

The international application of these techniques will depend on the availability of equivalent texts to the Soil Records, containing detailed descriptions of the relationships between soil and geology and landscape. Assuming such availability and the existence of detailed soil parent material mapping for training and testing purposes, these techniques should be applicable internationally.

The techniques investigated in this research have been developed with a target scale of 1:50,000. Where less detailed target scales of 1:250,000 or 1:1,000,000 are required, many of the same principals will apply, but greater consideration will need to be applied to the issue of membership of the map units. Multiple parent materials are likely to be present at such coarse scales, and complex map units will result. Such generalised soil parent material maps will not be suitable for addressing the types of detailed applications (for example digital soil mapping at a local scale) for which parent material maps are required. Nevertheless, an incorporation of levels of confidence about the membership of such broad units will allow for the propagation of the knowledge of errors.

8.7 Evaluation of the combined methodology

This methodology demonstrated that sparser data sampling led to less valuable maps than had been created with the detailed 60 m in the data mining methodology. Nevertheless, it was shown that expert knowledge, appropriately applied to these sparse

samples, could improve results and create suitable parent material maps with less than 1% of the sample points used in the data mining methodology.

Beyond certain sampling densities (typically between 1000 and 2000 m spacing) it was shown that the inputs derived in part from data mining tended to perform worse than those created from expert knowledge on its own, as at such wide sample spacings, there were too few points to create a reliable sample of the parent materials. This approach, using a sample of, say, between 700 and 1000 m, would, in a real world context, be significantly more cost-effective than that of the data mining methodology, and produce maps of higher value than the expert knowledge methodology.

The tests for association introduce methods of assessing the usefulness of input layers prior to the time-consuming creation and refinement of model inputs, particularly those which involve expert knowledge inputs. This has allowed similar results to be achieved in significantly less time. In research where more than three evidence layers are considered these approaches may be even more valuable as indicators of possible layers to use or remove.

The removal of the DTM pixel based SLOPE component has allowed a polygon vector based map to be created instead of the point based files in the previous methodologies. This not only produced more attractive maps, but also produces smaller file sizes and required significantly less time for the computer to process and plot the resulting data. Such issues may be considered trivial from a purely scientific viewpoint. However, from personal experience assisting end-users of LandIS soils data, it is known that the datasets which are simple to use are those which are most extensively used to address a wide range of environmental situations. Thus the ease of use of resulting datasets is also an important consideration. Should future methods require the input of point or pixel based data, it may be advisable to create polygon layers from such layers, if possible.

The expert knowledge methodology demonstrated that knowledge of the parent material classes within the study area can greatly increase the success of the initial prediction of parent material. This methodology has shown that in order to accurately characterise the

likely extent of parent material classes within a region, a systematic sampling is more effective than a clustered sampling, which is more typical of soil survey, due to land access issues and the need to understand soil / geology / landscape relationships. Furthermore, a systematic sampling has been shown to be more effective than basing the expected membership upon the composition of the map units of the National Soil Map (Table 39).

The combined methodology used sparse samplings to provide a source of quantitative data to supplement the expert knowledge gleaned from the published literature. A 700 m sample spacing was sufficient to predict parent material extents to within 1% on most parent materials in the Yeovil area (Table 39). A sample spacing of 1 km has been demonstrated to provide almost the same level of success of that of a 700 m grid sample (Figure 54). However, samples more widely spaced than 1000 m were not as effective at characterising the extent of the parent materials present in the area.

Some units of limited extent were missed in the sparse samplings. Nevertheless, as qualitative expert knowledge was available for the areas, these parent materials could also be included in the model, and were given appropriately low prior probabilities. In the case of chalk (BiB2), because of the small extent of this unit in the Yeovil area, it was missed by even the 700 m sampling strategy. But, as this sampling was being used to supplement and provide a quantitative framework for expert knowledge, which had previously identified the chalk parent material, this unit was well predicted ($\xi = 0.65$) by this methodology, and this additional class contributed to a higher map value.

The addition of expert knowledge to the data from the sparser samplings tended to increase the resulting map values, particularly in the more complex areas of Needwood Forest and Yeovil. The prediction of parent material in the relatively simple Worksop area tended not to be helped by the addition of expert knowledge to model inputs derived from sparse samples.

Beyond sample spacings of 2000 m (approximately 25 sample per 100km^2), the inputs derived, at least in part, from quantitative data mining tended to not produce maps as of

high value as those produced using just the qualitative expert knowledge. With so few samples, characterisation of parent material/covariate relationships can be inaccurate.

It has been demonstrated that more extensive parent material units tend to achieve higher class values (ξ). Nevertheless, distinctive parent material units can be identified from existing geological and soil mapping, even if they are not very extensive. Distinctive units include peat or chalk, which have obviously recognisable features, or alluvium, where there is a strong physical expression. This leads to closer agreement in linework between evidence layers such as geological maps and parent material mapping.

Units which are poorly predicted, relative to their extent, tend to result from different mapping priorities and depths between the soil and geological surveys. A key example is that of the confusion between the two limestone units (BhB1 and BhB2) where the differentiating characteristics are at a shallow depth which is not important for the creation of standard geological maps. Furthermore, the 1:250,000 scale National Soil Map (SOIL) is necessarily generalised and does not contain enough linework to differentiate between such similar classes at a nominal target scale of 1:50,000. Because of such scale differences, smaller units may also be omitted from less detailed evidence layers.

Because of these differences, the evidence layer datasets do not always have the required level of linework or classification detail to accurately predict every class of the NSRI parent material classification. Commonly, there is confusion between units which vary within the top 30 cm of the soil profile. In such situations, class amalgamation is warranted, and the presentation of the probabilities of all parent materials is encouraged.

Key points:

- Testing the association of evidence layers with the reference map can save time by highlighting potentially good or poor predictors.
- A systematic grid sample is ideal for the characterisation of parent material proportions within a study area.
- Model inputs created from sparse data samples between 700 and 1000 m can achieve similar results to denser sample grids, at a fraction of the cost.
- Expert knowledge can improve inputs derived from data mining at sparser sample densities.
- Samplings with spacings beyond approximately 2000 m add little to expert knowledge inputs.
- More extensive units tend to be better predicted, while distinctive units tend to be well predicted, irrespective of their extent.

9 CONCLUSIONS AND RECOMMENDATIONS

This chapter presents the conclusions of this research with discussion of the specific research objectives. Recommendations are provided for the implementation of parent material map creation using similar techniques, and the most effective use of the resulting maps. Areas which could benefit from future work are also identified. It is concluded that soil parent material maps may be derived from existing sources of information, albeit with certain limitations in parent material class detail, and with the understanding that some class confusion is inevitable. Nevertheless, the use of probabilities can convey some of the class uncertainty to the end users.

Four methodologies for the creation of soil parent material maps were developed, investigated and evaluated as part of this research. The data dictionary method used one-to-one translations of geological maps to parent material maps. The expert knowledge methodology extracted qualitative knowledge from published literature. This was used to define the relationships between parent material and three environmental covariates: geology, slope and a national soil map. These relationships were formalised into inputs to a corrected probability model which output the probability of the occurrence of each parent material class, given the environmental covariates. The data mining method mirrored that of the expert knowledge method, but derived its model inputs from extensive quantitative pairwise sampling on a 60 m grid across the study areas. The combined methodology sought a pragmatic way forward. It incorporated aspects of both expert knowledge and data from sparser sample grids, to create quantified expert knowledge model inputs. The nine research objectives are now discussed, after which recommendations are provided as to the effective application and implementation of the results of the research.

9.1 The identification of valuable soil parent material maps

The use of innovative metrics of map value, which accurately describe the multiple desired features of a map, can help when selecting the most effective techniques for the production of maps.

The value of a map depends on how well it fulfils the requirements of its users. It was desired that the parent material maps resulting from this research could help address a range of environmental issues. Specifically, in the context of this research, a valuable parent material map was defined as having numerous, clearly defined and highly specific parent material classes which are related to soil types. In addition the map would have a high overall accuracy, indicative of geographic accuracy.

With the clear definition of these attributes of valuable maps, a number of novel metrics were developed in order to quantitatively compare the results of the many maps which were produced in this research. Of particular note are the class value metric (ξ) and map value metric (ψ_3). The class value metric measures the spatial accuracy of particular units, calculating the geometric mean of the producer and user accuracy. The map value metric allows easy comparison of multiple desirable factors including the number, detail and accuracy of individual classes, as well as the overall map accuracy.

9.2 International and national classifications

The national NSRI classification offers the possibility of more detailed descriptions of the soil parent material than the international and lithological ESB classification as it is designed for use at a more detailed scale and contains parent material classes more closely related to soil.

Parent material maps were created using both international and national parent material classifications, from the European Soil Bureau (ESB) and the National Soil Resources Institute (NSRI). While the international ESB classification provided scope for

consistency across Europe, it was demonstrated that this classification produced less detailed and less accurate maps with fewer classes and lower map values (ψ_3) than equivalent maps using the lithological component of the national NSRI classification.

The ESB classification is closer to a lithological reinterpretation of a geological map than it is to a true soil parent material map, in that it does not address the physical nature of the parent material, but focuses entirely on the lithology of the material. Parent material maps should ideally make reference to aspects of both geology and soil, providing information on this transition zone, as does the NSRI classification. Furthermore, as many of the requirements such as for parent material maps (as described on page 4) are for application at a local scale, and so do not necessitate classification consistency across Europe, it was concluded that the NSRI classification should be used as the primary classification for maps in England and Wales. Should a consistent parent material map be required for wider international projects, the more detailed national classification could be converted to the international classification as required by means of translation tables or additional attribution.

While the national NSRI classification produces more valuable maps than the strongly lithological ESB classification, some class confusion remains. The detail required to differentiate between certain similar parent material classes of the NSRI classification is unlikely to be obtained from the environmental datasets which were used. This results from a lack of detail in some of the evidence layer maps, which were all at a coarser scale than the reference parent material maps, and also in the different priorities between the geological and soil survey organisations. For example, while soil surveyors note the stoniness of the top 30 cm of the soil column, this is rarely recorded on geological maps method.

9.3 Bedrock and surface geology layers

Surface geology maps, which include both superficial and bedrock components, provide a basis for more accurate predictions of soil parent material than bedrock-only maps, particularly in areas where surface deposits are extensive.

In the data dictionary methodology, surface geology maps (which consider both superficial and bedrock geology) and bedrock-only maps were compared to determine which produced the maps with the highest value (ψ_3). It was demonstrated that, while the superficial deposits may be incompletely mapped, the inclusion of these deposits in the geological input led to consistently higher map values in the areas where soils in drift were abundant. In the Worksop area, where there is little drift, the bedrock layer produced only very marginally higher map values. Therefore, the surface geology layer was used as a matter of course in all subsequent methodologies.

9.4 Class confusion and classification simplification

As a means of overcoming misclassifications, the amalgamation of commonly misclassified classes outperforms the use of simplified classifications. This is because it allows for class detail to be retained, wherever possible, only simplifying where necessary.

In the data dictionary method, extensive misclassification occurred when geological classes were translated to parent material classes. To overcome these misclassifications, two different methods of simplifying the parent material classifications were explored. The first simplified the entire classification on the basis of lithological similarity. The second amalgamated only the consistently misclassified units. It was found that parent material class amalgamation on a case by case basis resulted in maps of higher value than a simplification of the entire classification. The amalgamation approach retained as much classification detail as possible, only simplifying where necessary.

Some of the class confusion arose from the limitations of one-to-one translations as used in the data dictionary methodology. In the later methodologies, when expert knowledge and data mining were used to predict the probabilities of parent materials given environmental evidence layers, the models were more successful at the initial prediction of the most likely soil parent material. Class amalgamation continued to improve the resulting map values. However, because the initial predictions of parent material were better in these later methodologies, amalgamation did not bring about the same level of improvement as had been seen in the data dictionary method.

Because the later methodologies also include probabilities of occurrence for each of the parent material classes, it was suggested that an appropriate additional mechanism would be to consider each parent material map unit as a class with a defined probability distribution, dependant on the evidence layers.

Map value was often increased when units, predicted by evidence datasets lacking sufficient detail for class differentiation, were amalgamated. Class amalgamation also remains of use as a cartographic tool for representing heterogeneous units.

9.5 The use of expert knowledge to predict parent material

Expert knowledge, captured in published literature, was identified, extracted, and shown to be an able predictor of soil parent material, when combined with appropriate spatial datasets.

A novel approach to building parent material models was developed, making use of expert knowledge from published literature. This knowledge was identified, extracted and formalised into probabilistic model inputs. Extensive information was found in the Soil Records, (the books which accompany detailed soil maps) and other sources, including national soil and geological databases.

Much of the expert knowledge was of little use for predictive mapping, resulting from inconsistent levels of descriptive detail. Nevertheless, sufficient expert knowledge of the relationships between parent material and three environmental covariates (soil, slope and geology), was identified in order to develop inputs for a corrected and modified probability model. This model allowed the combination of probabilities from multiple evidence layers.

Additional map layers were identified to provide map evidence for the identified slope and soil related expert knowledge. The use of expert knowledge allowed a more detailed NSRI parent material classification to be used than in the data dictionary methodology, as the knowledge of parent material classes likely to be present in the areas reduced the number of potential classes.

Given the map evidence, the probability of occurrence of each parent material class was output from the model. The use of probabilities and multiple evidence layers removed the limitations of a one-to-one translation which hampered predictions in the data dictionary method and led to some misclassifications. The use of expert knowledge improved initial parent material prediction, and the additional evidence layers provided flexibility allowing the production of maps of higher value (ψ_3) than had previously been achieved.

Two key limitations of this method were identified. Firstly, considerable time was required to extract the expert knowledge from the literature and formalise it into model inputs, which had no guaranteed benefits. Indeed, it was shown that the slope input added little value to maps produced using just the geology and soil maps, and often had a negative effect. Secondly, the expert knowledge lacked quantitative data describing the relationships between parent material and the environmental covariates. The result was that most inputs to the probability model were constructed purely on the basis of qualitative descriptions. This gave rise to some relative over or under predictions of the geographic extent of parent material units. These limitations were addressed in later methodologies.

9.6 A quantitative data mining approach

Quantitative, pairwise sample data collected at a sample spacing of less than 1500 m produced maps of higher or comparable value to those produced by the use of expert knowledge. Denser sample strategies tended to produce maps of higher value.

In an attempt to quantify the effect that reliance on qualitative descriptions had on map value, a fully quantitative approach using the same model and map evidence layers used in the expert knowledge method was developed. In the data mining method, extensive sampling on a 60 m grid was carried out across the study areas to quantitatively define the pairwise relationships between the parent material and the evidence layers. This resulted in the production of parent material maps of high value.

The use of data mining techniques certainly increased the value of the parent material maps. However, concerns remained about the applicability of this approach in a previously unmapped area, as it was shown that considerably lower map values were achieved when models were trained on one area and then tested on a different area.

The effect on map value was considered when sample spacing was increased from 60 m to 700, 1400, 2100 and 2800 m. Such increases led to considerable decreases in map value, particularly in the more complex areas, where lower numbers of sample points could not quantify as effectively the relationships between parent material and the evidence layers. Commonly, parent material units of limited extent were unpredicted by the sparser samples. In the more complex study areas, at sample densities beyond 1500 m, the quantitative data mining techniques produced maps of lower value than were produced by the qualitative expert knowledge method.

9.7 A combined, quantified, expert approach

The combination of two inputs; formalised expert knowledge and sparse data sampling, can produce maps of higher value than either input individually, particularly in areas with greater geo-diversity.

The final methodology sought a pragmatic combination of aspects of the previous methods for application in a context more akin to the real world. The expert knowledge inputs from the second method were refined with the sparse data mining samples. This allowed the expert knowledge to be applied in a more quantitatively robust manner, while maintaining the knowledge of the units with limited extent.

The issue of the time needed to develop the expert knowledge inputs was addressed by calculating the association between the reference parent material map and each of the evidence layers. Evidence layers with limited association to parent material could be identified at an earlier stage, and expert knowledge not compiled for such predictors. While such approaches need to be undertaken with caution to prevent exclusion of potentially useful information, it was shown that the level of association between the slope map and parent material map was negligible compared to that of the geology or soil layer. On this basis, the slope evidence layer was removed from the model.

The combined methodology tended to produce maps of higher value than pure data mining at the sparser sample densities, particularly in the more complex areas. The inclusion of expert knowledge allowed minor parent material classes to be included in the models which were missed by the sparse data sampling.

Results do vary with the quality of the expert knowledge and the complexity of the area. This is demonstrated in the geologically simple Worksop area where the addition of expert knowledge rarely improved on the map value achieved by pure data mining. In the future, simple comparisons of the success of the maps produced by the pure expert knowledge method and those by sparse data sampling could be compared. These could

then be used to assess the appropriate level of reliance on the expert knowledge and data mined inputs for each of the parent material classes.

Further work is required on the extrapolation of expert knowledge from one area to an adjacent one. Success of such extrapolation is likely to be highly dependant upon the extent to which similar soil and geological units are found beyond the study area. Also important will be the consistency of mapping and the homogeneity of geological units beyond the study area. If the geological units give rise to similar parent materials and are mapped consistently, extrapolation is likely to be successful.

9.8 Parent material map fitness for purpose

The final maps produced in this research achieved relatively high overall accuracies and a reasonable number of well predicted and detailed effective parent material classes. Also provided was an assessment of the probability of prediction for each parent material type. Given these features it is concluded that maps produced by these methodologies are fit to address a range of environmental issues.

This research hypothesised that, with appropriate techniques, maps of soil parent material may be derived effectively from existing sources of information. The resulting parent material maps and classes were analysed for fitness for purpose according to the desirable attributes defined at the start of this research. A valuable parent material map would be very accurate, and have numerous, clearly defined and highly specific parent material classes which are related to soils types.

The national NSRI classification was shown to contain more detailed classes with closer links to soil than the international ESB classification. It was also demonstrated that some parent material units achieve higher class values (ξ) than others. Generally, this was shown to be strongly dependant on the extent of the parent material class in the study area. However, certain parent material units were shown to be more easily predicted than would be indicated by their limited extent. These included chalk, peat

and alluvium, which are distinctive in terms of chemistry or physical structure or are closely defined by the landscape.

It was shown that the available geological, soil and slope evidence layers lacked the necessary detail in linework and attribution to accurately predict certain parent materials. These include thin drift or those with differentiating characteristics in the top 45 cm of the soil profile. The inability to predict all parent material types is a limitation of the resulting maps. This arises from three main issues: the different mapping priorities between the geological and soil survey organisations; the lack of detail in the 1:250,000 scale soil map; and the distribution of parent material types across many slope classes.

Nevertheless, overall accuracies of the most likely parent material range between 65% and 90%, while still maintaining a useful number of parent material classes. Such levels are equivalent to the assessments of soil map unit heterogeneity by Sturdy (1971). Furthermore, it has been shown that where the most likely parent material did not agree with the reference parent material map, it was common that the second or third most probable parent material was in agreement and that the difference in the probabilities could be very small. This understanding of the probability of prediction is useful for input into a range of environmental models, and represents an improvement on traditional map reinterpretations (similar to those in the data dictionary methodology) with pure units and no measure of confidence.

The parent material maps created with the combined, quantified expert knowledge methodology had high levels of overall agreement. These maps included indications of the probability of predictive success and contained acceptable numbers of detailed classes which are closely associated with soil type. Therefore, this approach appears likely to offer a pragmatic approach to the creation of parent material maps which are able to help address a range of environmental issues.

9.9 Contributions to knowledge

A number of new findings have been made in this research, some of which deserve particular mention as contributions to knowledge. These include:

Novel or improved method, models and metrics

- The use of expert knowledge extracted from literature and formalised to derive inputs for probability models for the mapping of parent material.
- The combination of expert knowledge with sparse data mining to predict parent material.
- The corrected and expanded probability model.
- The new map value (ψ_3) and class value (ξ) metric which have been useful in providing a more holistic assessment of the success of models inputs.

Improved understanding

- That the national NSRI parent material classification is more appropriate for detailed parent material maps than the international ESB classification.
- That class amalgamation on a case by case basis is a more appropriate way of simplifying parent material classifications than lithological simplification of the entire classification.
- That surface geology tends to be a better predictor of parent material than bedrock geology, particularly in areas with extensive drift.
- That slope does not predict parent material as well as 1:50,000 scale geology maps or 1:250,000 scale soils maps in the Worksop, Needwood Forest and Yeovil study areas, which are all lowland regions.
- That at very sparse sample densities, expert knowledge tends to be a better predictor of parent material than quantitative sampling.
- That inputs derived from quantitative sampling can be improved with expert knowledge derived from literature, particularly in areas with high geo-diversity.

9.10 Recommendations

On the basis on this research, some recommendations are provided for the creation of parent material maps at a nominal 1:50,000 scale from existing environmental datasets. Recommendations are also provided for appropriate usage of these maps, and for future work.

Use a probability model to create the parent material map rather than a simple one-to-one translation

A probability model can be used to provide probabilities of each parent material class given a range of evidence inputs. The use of this type of model is recommended instead of a simple one-to-one translation of an existing map as it allows the integration of multiple sources of information and also provides a statement of the trustworthiness or confidence of the resulting maps. These are useful for the transfer of knowledge of error to later research applications.

Undertake a systematic grid sampling to characterise parent materials and their relationships with the evidence layers

Ideally, a systematic sampling of the parent materials and the evidence layers with a grid spacing less than 1000 m, should be carried out, in conjunction with expert knowledge, in order to characterise the likely parent material classes and their approximate proportions. These proportions can be entered into the model as the prior probabilities for each parent material class. Alternatively, wider sample spacing may be employed to characterise the likely parent material extents within an area. However, samples with wider spacings tend to miss more map classes and provide less accurate predictions of class extents. In such cases, a stronger reliance on expert knowledge will be required to supplement the list of parent material classes likely to be present in the area.

Calculate association between parent material and the evidence layers

There can be significant work entailed in the process of building model input layers, particularly those derived from expert knowledge. It is therefore recommended that the association between evidence layers and the reference map or reference sample points should be tested prior to the creation of model inputs for these evidence layers. This can save considerable time by not deriving expert knowledge inputs for less useful inputs. Such tests can be carried out on an adjacent training area with similar a landscape to the desired study area and a detailed map, or by sparse sampling across the area in question.

Use the NSRI parent material classification for maps in England and Wales

As well as being lithological, the NSRI parent material classification uses classes more closely linked to defined soil series in England and Wales. As such it addresses both geology and soil themes and can assist in supporting a range of environmental models, as well as being a useful input into soil models. If international harmonisation is required, this can be achieved with additional attribution or by means of an appropriate look-up table.

Use surface geology rather than bedrock geology

Whilst the extent of superficial geology mapped by the geological survey is less than that mapped by the soil survey, it has been demonstrated that maps using the surface geology outperform bedrock-only inputs where drift is abundant. These should therefore be used in preference to bedrock-only inputs.

Extract expert knowledge from literature, if available

Expert knowledge extracted from published literature and formalised has been shown to be capable of producing useful model inputs, both on its own and in conjunction with quantitative data. This has been an important finding of this research. If time permits, expert knowledge should be assessed for study areas. If this knowledge is found to be consistent and sufficiently detailed, it can be used to derive new model inputs, or to verify or modify existing ones.

Class membership should be considered in terms of probabilities

The parent material map classes produced by this method can be considered to have complex membership. These memberships can be based either on the probability distributions resulting from the models, or from the relative membership gleaned from the sparse data mining sample after running the model, combined with expert knowledge to provide supplementary information for units with smaller extents.

Use guided amalgamations rather than simplifying the whole classification

Class amalgamation can be used to group commonly misclassified 'most likely' units. Amalgamations are useful as cartographic tools and for presenting 'mixed units' where there is exists a higher level of uncertainty. But for future environmental models using the parent material map, the probability of each parent material class is likely to be more useful than just a single parent material class. Therefore, even when classes are amalgamated, the probability of each parent material class should still be associated with the map.

9.11 Future work

A number of areas which could benefit from addition research have been identified. These are now briefly discussed.

Extrapolation and additional study areas

The application to adjacent or similar areas of model inputs based on expert knowledge, extracted from published literature regarding specific map sheets, would benefit from further investigation. This would likely involve the characterisation of similar areas on the basis of similar geological and soil units. Additional detailed soil maps in these areas, from which reference parent material maps could be derived, would need to be identified for test purposes.

The National Soil Map evidence layer did not perform as well in the Yeovil area as in the Needwood Forest and Worksop areas, which used the detailed mapping as the basis for the National Map. Additional soil map sheets created after the mapping of the National Soil Map should be identified to further test the usefulness of this layer as a predictor of parent material.

The study areas in this research were dominated by sedimentary rocks and quaternary deposits. These methodologies should be tested on a range of additional geological landscapes, including peat-rich uplands, areas where the parent material is influenced by igneous or metamorphic rocks and coastal regions. Additional test areas should be selected so all the major parent material classes are tested.

Development of metrics and models

It would be beneficial to incorporate a measure of taxonomic distance in the class and map value metrics. This would mean that, when amalgamated, parent materials which are very similar would not be penalised as severely as very different classes. A

difficulty would be in determining the basis for class similarity, be it structural, lithological or both, but this would increase the usefulness of the map and class value metrics.

Before deriving combined inputs from expert knowledge and sparse data samples, the maps produced by pure expert knowledge and the pure sparse samples should be compared. An assessment could then be made of the appropriate level of reliance on the expert knowledge and data mined inputs for each of the parent material classes.

This research examined pairwise data mining in comparison with expert knowledge. Additional data layers offer the possibility of refining linework and classification detail. Additional layers may be identified using the described tests for association with parent material. Following identification, structured interviews with experts could be used to populate or refine model inputs.

The use of the new BGS parent material map (Lawley, 2009) should be investigated as a more sophisticated geological input into parent material models. Of particular interest are the fields describing the gravel deposits, and whether the extent of superficial deposits now more closely matches the extent shown by soil maps.

Multivariate data mining (with many more landscape attributes, remote sensing layers, and geographic datasets) should be investigated as a modelling technique in its own right. The results of such techniques could then be incorporated with the methods discussed in this research.

This research used a probability model to predict parent material. One convenient assumption in this model was that of conditional independence, which does not always hold true. Methods of relaxing this assumption should be investigated. Alternative approaches such as multinomial logistic regression, neural nets or decision trees should also be investigated to see if improved results can be achieved using these methods.

The new technique for conveying the map purity of the evidence layers, proposed in Farewell and Farewell (2010), should be investigated further in additional study areas and compared with the defined map purity approach used in this research.

Research into the most effective way of characterising class membership would be beneficial. Two options discussed include the definition of class membership on the basis of probabilities or by sampling of a known area. These two approaches need to be compared and reported on.

REFERENCES

Agbenin, J. O. and Tiessen, H. (1995), "Soil Properties and their variation on two contiuous hillslopes in Northeast Brazil", *Catena,* vol. 24, pp. 147-161.

Agbu, P. A. and Olson, K. R. (1992), "Model to predict soil parent material underlying a loess mantle in Illinois from satellite data", *Soil Science,* vol. 153, no. 2, pp. 142-148.

Avery, B. W. and Soil Survey of England and Wales. (1987), *Soil Survey laboratory methods,* Soil Survey of England and Wales Rothamsted Experimental Station, Harpenden.

Bellamy, P. H., Loveland, P. J., Bradley, R. I., Lark, R. M. and Kirk, G. J. D. (2005), "Carbon losses from all soils across England and Wales 1978–2003", *Nature,* vol. 437, no. 8, pp. 245 - 248.

Benson, A. K. and Hash, S. T. (1998), "Integrated three-dimensional interpretation of major concealed faults beneath Mapleton, Utah County, Utah using gravity data, supported with magnetic data", *Engineering Geology,* vol. 51, no. 2, pp. 109-130.

Berger, J. O. and Jefferys, W. H. (1991), *The application of robust Bayesian analysis to hypothesis testing and Occam's Razor,* # 91-04, Purdue University, West Lafayette.

BGR (2004), *Soil Regions of the European Union and Adjacent Countries 1:5,000,000*, Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover and Berlin.

BGS (2010), *Geosure*, British Geological Survey, Keyworth.

Billett, M. F., Lowe, J. A. H., Black, K. E. and Cresser, M. S. (1997), "The influence of parent material on small-scale spatial changes in streamwater chemistry in Scottish upland catchments", *Journal of Hydrology,* vol. 187, no. 3-4, pp. 311-331.

Bishop, K. H., Grip, H. and O'Neill, A. (1990), "The origins of acid runoff in a hillslope during storm events", *Journal of Hydrology,* vol. 116, no. 1-4, pp. 35-61.

Blum, W. E. H. (1993), "Soil Protection concept of the Council of Europe and Integrated Soil Research", in Eijsackers, H. and Hamers, T. (eds.) *Soil and Environment Vol I - Integrated soil and sediment research: a basis for proper protection,* Kluwer Academic Publisher, Dordrecht.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures", *Statistical Science,* vol. 16, no. 3, pp. 299-231.

British Geological Survey (2009), *BGS Lexicon*, available at: http://www.bgs.ac.uk/lexicon/home.cfm (accessed January 19).

British Geological Survey, (2007), *Digital Geological Map of Great Britain 1:50,000 scale (DiGMapGB-50)*, Version 2.11 ed., British Geological Survey, Keyworth, Nottingham.

Bui, E. N. (2004), "Soil survey as a knowledge system", *Geoderma,* vol. 120, no. 1-2, pp. 17-26.

Bui, E. N., Henderson, B. L. and Viergever, K. (2006), "Knowledge discovery from models of soil properties developed through data mining", *Ecol.Model.,* vol. 191, no. 3-4, pp. 431-446.

Bui, E. N. and Moran, C. J. (2003), "A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia", *Geoderma,* vol. 111, no. 1-2, pp. 21-44.

Bui, E. N. and Moran, C. J. (2001), "Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data", *Geoderma,* vol. 103, no. 1-2, pp. 79-94.

Bui, E., Loughhead, A. and Corner, R. (1999), "Extracting soil-landscape rules from previous soil surveys", *Aust.J.Soil Res.,* vol. 37, no. 3, pp. 495-508.

Caspari, T., Bäumler, R., Norbu, C., Tshering, K. and Baillie, I. (2006), "Geochemical investigation of soils developed in different lithologies in Bhutan, Eastern Himalayas", *Geoderma,* vol. 136, no. 1-2, pp. 436-458.

Cauvin-Cayet, C., Galdeano, A., Egal, E., Pozzi, J. P. and Truffert, C. (2001), "Magnetic Modellin in the French Cadomian belt (northern Armorican Massif)", *Tectonophysics,* vol. 331, pp. 123-144.

Chatfield, C. (1995), "Model uncertainty, data mining and statistical inference.", *Journal of the Royal Statistical Society,* vol. 158, no. 3, pp. 419-466.

Chatterjee, R. S. (2003), "Structural pattern of Holenarsipur Supracrustal Belt, Karnataka, India as observed from digitally enhanced high resolution multi-sensor optical remote sensing data aided by field survey", *International Journal of Applied Earth Observation and Geoinformation,* vol. 4, pp. 195-215.

Clayden, B. and Hollis, J. M. (1984), *Criteria for differentiation soil series,* SSEW, Harpenden.

Cohen, J. (1960), "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement,* vol. 20, no. 1, pp. 37-46.

Colborne, G. J. N. and Staines, S. J. (1987), *Soils in Somerset I - Sheet St41/51 (Yeovil),* Harpenden.

Cook, S. E., Corner, R. J., Groves, P. R. and Grealish, G. J. (1996), "Use of airborne gamma radiometric data for soil mapping", *Aust.J.Soil Res.,* vol. 34, pp. 183-194.

Corner, R. J., Hickey, R. J. and Cook, S. E. (2002), "Knowledge Based Soil Attribute Mapping In GIS: The Expector Method", *Transactions in GIS,* vol. 6, no. 4, pp. 383-402.

Dent, D. (2007), "Environmental geophysics mapping salinity and water resources", *International Journal of Applied Earth Observation and Geoinformation,* vol. 9, no. 2, pp. 130-136.

Dickson, B. L. and Scott, K. M. (1990), *Radioelement distributions in wethered granitoids and aeolian soils in NSW. AMIRA P263: Improving the interpretation of airborne gamma-ray surveys. CSIRO Division of Exploration Geoscience, .*

Dickson, B. L. (2004), "Recent advances in aerial gamma-ray surveying", *J.Environ.Radioact.,* vol. 76, pp. 225-236.

Dobos, E., Carre, F., Hengl, T., Reuter, H. I. and Toth, G. (2006), *Digital Soil Mapping as a support to production of functional maps EUR 22123,* , Office for Official Publications of the European Communities, Luxemburg.

Dowling, J. W. F. (1966), "The mode of occurrence of laterites in northern Nigeria and their appearance in aerial photography", *Engineering Geology,* vol. 1, no. 3, pp. 221-233.

Doyle, P. J. and Fletcher, W. K. (1977), *Canadian Journal of Plant Science,* vol. 57, pp. 859-864.

Duda, R. O., Hart, P. E., Barrett, P., Gasching, J. G., Kilge, K. and Reboh, R. S.,J. (1978), *Development of the Prospector Consultation System for Mineral Exploration,* (Final Report for SDI Projects 5821 and 6415), SRI International Artificial Intellegence Center, Stamford, CA.

e-SOTER (2008), *(e-SOTER) Regional pilot platform as EU contribution to a Global Soil Observing System. Annex 1 - Description of Work,* (Grant agreement no.: 211578).

European Soil Bureau, (2001), *Soil Geographical Database of Eurasia at scale 1:1,000,000 version 4 beta, 25/09/2001*, European Soil Bureau, Ispra, Italy.

FAO (1995), *Global and national soils and terrain digital databases (SOTER) Procedures Manual,* , Land and Water Development Division (FAO), Rome.

Farewell, T. S. and Farewell, D. M. (2010), "Knowledge-based soil attribute mapping in GIS: corrections and extensions to the Expector method", *Transactions in GIS,* vol. (in press).

Farmer, L. (2008), *Investigation of remote sensing and geographical information system techniques for increased versatility within landcover map production,* , PHD Thesis, Cranfield University, Cranfield.

Fealy, R. M., Green, S., Loftus, M., Meehan, R., Radford, T., Cronin, C. and Bulfin, M. (2009), *Teagasc EPA Soil and Subsoils Mapping Project-Final Report. (Unpublished),* , Teagasc/EPA, Dublin.

Feng, Z. and Johnson, W. C. (1995), "Factors affecting the magnetic susceptibility of a loess-soil sequence, Barton County, Kansas, USA", *Catena,* vol. 24, no. 1, pp. 25-37.

Findlay, D. C. (1970), *Making 1:25,000 soil maps,* , Westbury-on-Trym, Bristol.

Findlay, D. C. and Soil Survey of England and Wales (1984), *Soils and their use in South West England,* Soil Survey of England and Wales, Harpenden.

Finke, P.A., Hartwich, R., Dudal, R., Ibàñez, J.J., Jamagne, M., King, D., Montanarella, L. and Yassoglou, N. (1998). Georeferenced Soil Database for Europe, Manual of Procedures, Version 1.0. - European Soil Bureau Research report No. 5, Office for Official Publications of the EC (EUR 18092 EN), Luxembourg. 184 pp.

Freeland, R. S., Yoder, R. E. and Ammons, J. T. (1998), "Mapping shallow underground features that influence site-specific agricultural production", *Journal of Applied Geophysics,* vol. 40, no. 1-3, pp. 19-27.

Frey, B., Kremer, J., Rüdt, A., Sciacca, S., Matthies, D. and Lüscher, P. (2009), "Compaction of forest soils with heavy logging machinery affects soil bacterial community structure", *European Journal of Soil Biology,* vol. 45, no. 4, pp. 312-320.

Galdeano, A., Asfirane, F., Truffert, C., Egal, E. and Debeglia, N. (2001), "The aeromagnetic map of the French Cadomian belt", *Tectonophysics,* vol. 331, pp. 99-122.

Gerber, R., Salat, C., Junge, A. and Felix-Henningsen, P. (2007), "GPR-based detection of Pleistocene periglacial slope deposits at a shallow-depth test site", *Geoderma,* vol. 139, no. 3-4, pp. 346-356.

Gomez Valle, R., Friedman, J. D., Gawarecki, S. J. and Banwell, C. J. (1970), "Photogeologic and thermal infrared reconnaissance surveys of the Los Negritos-Ixtlan de los Hervores geothermal area, Michoacan, Mexico", *Geothermics,* vol. 2, no. Part 1, pp. 381-388, 389-396, IN5-IN6, 397-398.

Gomez, C., Delacourt, C., Allemand, P., Ledru, P. and Wackerle, R. (2005), "Using ASTER remote sensing data set for geological mapping, in Namibia", *Physics and Chemistry of the Earth, Parts A/B/C,* vol. 30, no. 1-3, pp. 97-108.

Goodman, L. A. (1960), "On the exact variance of products", *Journal of the American Statistical Association,* vol. 55, no. 292, pp. 708-713.

Grieve, I. C. (1999), "Effects of parent material on the chemical composition of soil drainage waters", *Geoderma,* vol. 90, no. 1-2, pp. 49-64.

Grimley, D. A., Arruda, N. K. and Bramstedt, M. W. (2004), "Using magnetic susceptibility to facilitate more rapid, reproducible and precise delineation of hydric soils in the midwestern USA", *Catena,* vol. 58, no. 2, pp. 183-213.

Hannam, J. A. and Dearing, J. A. (2008), "Mapping soil magnetic properties in Bosnia and Herzegovina for landmine clearance operations", *Earth and Planetary Science Letters,* vol. 274, no. 3-4, pp. 285-294.

Hansen, M. K., Brown, D. J., Dennison, P. E., Graves, S. A. and Bricklemyer, R. S. (2009), "Inductively mapping expert-derived soil-landscape units within dambo wetland catenae using multispectral and topographic data", *Geoderma,* vol. 150, no. 1-2, pp. 72-84.

Hewson, R. D., Cudahy, T. J., Mizuhiko, S., Ueda, K. and Mauger, A. J. (2005), "Seamless geological map generation using ASTER in the Broken Hill-Curnamona province of Australia", *Remote Sensing of Environment,* vol. 99, no. 1-2, pp. 159-172.

Hodgson, J. M. (1997), *Soil Survey Field Handbook. Soil Survey Technical Monograph No. 5,* , Silsoe.

Hollis, J. M. (1978), *Soils in Salop 1- Soil Survey Record No.49,* 1st ed, Soil Survey of England and Wales, Harpenden.

Homburg, J. (2005), "ARCHEOLOGY IN RELATION TO SOILS", in Daniel Hillel (ed.) *Encyclopedia of Soils in the Environment,* Elsevier, Oxford, pp. 95-102.

Igué, A. M., Gaiser, T. and Stahr, K. (2004), "A soil and terrain digital database (SOTER) for improved land use planning in Central Benin", *European Journal of Agronomy,* vol. 21, no. 1, pp. 41-52.

Intermap Technologies (2002), *NEXTMap Digital Terrain Model (5 m)*.

James, I. T., Waine, T. W., Bradley, R. I., Taylor, J. C. and Godwin, R. J. (2003), "Determination of Soil Type Boundaries using Electromagnetic Induction Scanning Techniques", *Biosystems Engineering,* vol. 86, no. 4, pp. 421-430.

Jarvis, M.G., Hazelden , J. and Mackeny, D. (1979) *Soils of Berkshire,* Soil Survey Bulliten No. 8, Harpenden

Jenny, H. (1941), *Factors of soil formation. A system of quantitative Pedology,* McGraw-Hill, New York.

Jones, R. J. A. (1983), *Soils in Staffordshire III - Sheet SK02/12 (Needwood Forest),* , Harpenden.

Jones, R. (2006), *Personal Communication,* .

Katz, S. S. (1991), "Emulating the prospector expert system with a raster gis", *Computers & Geosciences,* vol. 17, no. 7, pp. 1033-1050.

Keay, C.A., Hallett, S.H., Farewell, T.S., Rayner, A.P. & Jones, R.J.A. (2009) "Moving the National Soil Database for England and Wales (LandIS) towards INSPIRE Compliance", International Journal Of Spatial Data Infrastructures Research, 134-155.Kogan, R. M., Nazarov, I. M. and Fridman, S. D. (1969), "Gamma-ray spectrometry of natural environments and formations", *Israel Program for Scientific Translations,* .

Koons, R. D., Helmke, P. A. and Jackson, M. L. (1980), "Association of trace elements with iron oxides during rock weathering", *Soil Sci.Soc.Am.J.,* vol. 44, pp. 155-159.

Krol, B., Rossiter, D. G. and Siderius, W. (2004), "Ontologies for multi-source data integration in predictive soil mapping", in Lagacherie, P., McBratney, A. B. and Voltz, M. (eds.) *Digital Soil Mapping: An Introductory Perspective,* 1st ed, Elsevier, Amsterdam, pp. 121-135.

Lado, L. R., Hengl, T. and Reuter, H. I. (2008), "Heavy metals in European soils: A geostatistical analysis of the FOREGS Geochemical database", *Geoderma,* vol. 148, no. 2, pp. 189-199.

Laffan, S. W. and Lees, B. G. (2004), "Predicting regolith properties using environmental correlation: a comparison of spatially global and spatially local approaches", *Geoderma,* vol. 120, pp. 241-258.

Lagacherie, P. and Holmes, S. (1997), "Addressing geographical data errors in a classification tree for soil unit prediction", *International Journal of Geographical Information Science,* vol. 11, no. 2, pp. 183-198.

Lagacherie, P., Legros, J. P. and Burfough, P. A. (1995), "A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area", *Geoderma,* vol. 65, no. 3-4, pp. 283-301.

Lagacherie, P., Baret, F., Feret, J., Madeira Netto, J. and Robbez-Masson, J. M. (2008), "Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements", *Remote Sensing of Environment,* vol. 112, no. 3, pp. 825-835.

Lambert, J. J., Daroussin, J., Eimberck, M., Le Bas, C., Jamagne, M., King, D. and Montanarella, L. (2003), *Soil Geographical Database for Eurasia & the Mediterranean: Instruction Guide for Elaboration at scale 1:1,000,000 version 4.0,* EUR 20422 EN, European Soil Bureau, JRC, Ispra.

Lark, R. M., Bishop, T. F. A. and Webster, R. (2007), "Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties", *Geoderma,* vol. 138, no. 1-2, pp. 65-78.

Lavreau, J. and Fernandez-Alonso, M. (1991), "Correcting airborne raqdiometric data for water/vegetation screening using Landsat Thematic Mapper imagery", .

Lawley, R. (2009), *The soil-parent material database: A User Guide,* , British Geological Survey, Keyworth, Nottingham.

Lawley, R. and Smith, B. (2008), "Digital Soil Mapping at a National Scale: A Knowledge and GIS Based Approach to Improving Parent Material and Property Information", in Hartemink, A. E., McBratney, A. B. and Mendonca Santos, M. L. (eds.) *Digital Soil Mapping with Limited Data,* 1st ed, Springer, Netherllands, pp. 173-182.

Lorenz, K. and Lal, R. (2009), "Biogeochemical C and N cycles in urban soils", *Environment international,* vol. 35, no. 1, pp. 1-8.

Lundin, L. (2006), *Soil Parent Material*, available at: http://www-markinfo.slu.se/eng/soildes/jordart.html (accessed June 22nd 2009).

Macaulay, S. and Mullen, I. (2007), "Predicting salinity impacts of land-use change: Groundwater modelling with airborne electromagnetics and field data, SE

Queensland, Australia", *International Journal of Applied Earth Observation and Geoinformation,* vol. 9, no. 2, pp. 124-129.

Maes, S. M., Tikoff, B., Ferré, E. C., Brown, P. E. and Miller Jr., J. D. (2007), "The Sonju Lake layered intrusion, northeast Minnesota: Internal structure and emplacement history inferred from magnetic fabrics", *Precambrian Research,* vol. 157, no. 1-4, pp. 269-288.

Manta, D. S., Angelone, M., Bellanca, A., Neri, R. and Sprovieri, M. (2002), "Heavy metals in urban soils: a case study from the city of Palermo (Sicily), Italy", *The Science of the total environment,* vol. 300, no. 1-3, pp. 229-243.

Mayr, T. R., Palmer, R., Lawley, R. and Fletcher, P. (2001), *New Methods of Soil Mapping - Final Report for Defra Project SR0120,* , National Soil Resources Institute, Cranfield University, Silsoe.

McBratney, A. B., Mendonca Santos, M. L. and Minasny, B. (2003), "On digital soil mapping", *Geoderma,* vol. 117, pp. 3-52.

McDonald, P. A. and Pettifer, G. R. (1992), "The application of airborne radiometric classification techniques to salinity studies in Western Victoria", *Geological Survey of Australia, Abstracts,* vol. 32, pp. 305.

McKenzie, N. J. and Ryan, P. J. (1999), "Spatial prediction of soil properties using environmental correlation", *Geoderma,* vol. 89, pp. 67-94.

Minasny, B., McBratney, A. B. and Lark, R. M. (2008), "Digital Soil Mapping Technologies for Countries with Sparse Data Infrastructures", in Hartemink, A. E., McBratney, A. B. and Mendonca Santos, M. L. (eds.) *Digital Soil Mapping with Limited Data,* Springer Netherlands, , pp. 15-30.

Minasny, B. and McBratney, A. B. (2007), "Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes", *Geoderma,* vol. 142, no. 3-4, pp. 285-293.

Minty, B. R. S. (1996), "The fundamentals of airborne gamma-ray spectrometry", *AGSO Journal of Australian Geology and Geophysics,* vol. 17 (2).

Moghtaderi, A., Moore, F. and Mohammadzadeh, A. (2007), "The application of advanced space-borne thermal emission and reflection (ASTER) radiometer data in the detection of alteration in the Chadormalu paleocrater, Bafq region, Central Iran", *Journal of Asian Earth Sciences,* vol. 30, no. 2, pp. 238-252.

Moles, N. R. and Moles, R. T. (2002), "Influence of geology, glacial processes and land use on soil composition and Quaternary landscape evolution in The Burren National Park, Ireland", *Catena,* vol. 47, no. 4, pp. 291-321.

Moncoulon, D., Probst, A. and Party, J. (2004), "Weathering, atmospheric deposition and vegetation uptake: role for ecosystem sensitivity to acid deposition and critical load", *Comptes Rendus Geosciences,* vol. 336, no. 16, pp. 1417-1426.

Moran, C. J. and Bui, E. (2002), "Spatial data mining for enhanced soil map modelling", *International Journal of Geographical Information Science,* vol. 16, no. 6, pp. 533-549.

Morari, F., Castrignanò, A. and Pagliarin, C. (2009), "Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors", *Computers and Electronics in Agriculture,* vol. 68, no. 1, pp. 97-107.

Mosley, M. P. (1982), "Subsurface flow velocities through selected forest soils, South Island, New Zealand", *Journal of Hydrology,* vol. 55, no. 1-4, pp. 65-92.

Mullen, I. C., Wilkinson, K. E., Cresswell, R. G. and Kellett, J. (2007), "Three-dimensional mapping of salt stores in the Murray-Darling Basin, Australia: 2. Calculating landscape salt loads from airborne electromagnetic and laboratory data", *International Journal of Applied Earth Observation and Geoinformation,* vol. 9, no. 2, pp. 103-115.

Northey, R. D. (1974), "Insurance Claims from Earthquake Damage in Relation to Soil Pattern", in R.W. Simonson (ed.) *Developments in Soil Science,* Elsevier, , pp. 151-159.

NSRI (2008a), *The Digital National Soil Map - NATMAPvector*, Cranfield University, Cranfield.

NSRI (2008b), *The National Soil Inventory*, Cranfield University, Cranfield.

NSRI (2008c), *The Simplified National Soil Map - NATMAPsoilscapes*, Cranfield University, Cranfield.

NSRI (2009a), *LandIS Soil Attribute Data*, available at: http://www.landis.org.uk/gateway/ooi/series.cfm (accessed January 19, 2009).

NSRI, (2009b), *Natural Perils Directory*, Cranfield University.

Odeh, I. O. A. and McBratney, A. B. (2000), "Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of Eastern Australia", *Geoderma,* vol. 97, pp. 237-254.

Odell, G. and Lofgren, O. (2006), *A short summary of Swedish Survey of Forest Soils and Vegetation*, available at: http://www-sml.slu.se/sk/skeng.htm (accessed 22nd June 2009).

Olsson M (1999) Soil survey in Sweden. In: Bullock, P., Jones, R.J.A., Montanarella, L. (eds) Soil resources of Europe, the european soil bureau, Joint Research Centre, Ispra, Italy, pp 145-151

Palmer, R. C., Bellamy, P. H., Truckell, I. C., Farewell, T. S. and Cooke, H. J. (2007), *Review of accuracy, relevance and value of the BGS Soil Parent Material map and its associated data fields. NSRI research report No. YE20041E for BGS,* , NSRI, Cranfield University, Silsoe.

Peng, W., Wheeler, D. B., Bell, J. C. and Krusemark, M. G. (2003), "Delineating patterns of soil drainage class on bare soils using remote sensing analyses", *Geoderma,* vol. 115, no. 3-4, pp. 261-279.

Phillips, J. D. and Marion, D. A. (2005), "Biomechanical effects, lithological variations, and local pedodiversity in some forest soils of Arkansas", *Geoderma,* vol. 124, no. 1-2, pp. 73-89.

Plackett, R. L. (1983), "Karl Pearson and the Chi-squared Test", *International Statistical Review,* vol. 51, pp. 59-72.

Qi, F. and Zhu, A. -. (2003), "Knowledge discovery from soil maps using inductive learning", *International Journal of Geographical Information Science,* vol. 17, no. 8, pp. 771-795.

Qi, F., Zhu, A., Harrower, M. and Burt, J. E. (2006), "Fuzzy soil mapping based on prototype category theory", *Geoderma,* vol. 136, no. 3-4, pp. 774-787.

Ragg, J. M. and Soil Survey of England and Wales (1984), *Soils and their use in Midland and Western England,* Soil Survey of England and Wales.

Ramli, A. T. (1996), "Environmental Terrestrial Gamma Radiation Dose and its relationship with Soil Type ans Underlying Geological Formations in Pontian District, Malaysia", *Appl.Radiat.Isot.,* vol. 48, no. 3, pp. 407-412.

Reeve, M. J. (1976), *Soils in Nottinghamshire III - Sheet SK57 (Worksop),* , Harpenden.

RI USDA NRCS (2009), *Soil Parent Materials of Rhode Island*, RI USDA NRCS.

RIGIS (1989), *Rhode Island Glacial Deposits*, RIGIS, Providence, Rhode Island.

Rodríguez Martín, J. A., Arias, M. L. and Grau Corbí, J. M. (2006), "Heavy metals contents in agricultural topsoils in the Ebro basin (Spain). Application of the multivariate geoestatistical methods to study spatial variations", *Environmental Pollution,* vol. 144, no. 3, pp. 1001-1012.

Rossiter, D. G. (2005), "Digital soil mapping: Towards a multiple-use Soil Information System", *Análisis Geográficos (Revista del Instituto Geográfico "Augusín Codazzi"),* vol. 32, no. 1, pp. 7-15.

Roy, K., Allen, E., Barge, J., Ross, J., Curran, R., Bogucki, D., Franzi, D., Kretser, W., Frank, M., Spada, D. and Banta, J. (1997), *Influences on Wetlands and Lakes in the Adirondack Park of New York State: A Catalog of Existing and New GIS Data Layers for the 400,000 Hectare Oswegatchie/Black River Watershed,* CD992087-01, New York State Adirondack Park Agency, State University of New York at Plattsburgh, Adirondack Lakes Survey Corporation, New York.

Schetselaar, E. M., Chung, C. F. and Kim, K. E. (2000), "Integration of Landsat TM, Gamma-Ray, Magnetic and Field Data to Discriminate Lithological Units in Vegetated Granite-Gneiss Terrain", *Remote Sensing of the Environment,* vol. 71, pp. 89-105.

Scott, M. L. and Needelman, B. A. (2007), "Utilizing water well logs for soil parent material mapping in the mid-atlantic coastal plain", *Soil Science,* vol. 172, no. 9, pp. 701-720.

Scull, P., Franklin, J., Chadwick, O. A. and McArthur, D. (2003), "Predictive soil mapping: a review", *Progress in Physical Geography,* vol. 27, no. 2, pp. 171-197.

Sen, M. K., Stoffa, P. L., Seifoullaev, R. K. and Fokkema, J. T. (2003), "Numerical and Field Investigations of GPR: Toward an Airborne GPR", *Subsurface Sensing Technologies and Applications,* vol. 4, no. 1, pp. 41-60.

Shotton, F. W., Keen, D. H., Coope, G. R., Currant, A. P., Gibbard, P. L., Aalto, M., Peglar, S. M. and Robinson, J. E. (1993), "The middle pleistocene deposits of Waverley Wood Pit, Warwickshire, England", *Journal of Quaternary Science,* vol. 8, no. 4, pp. 293-325.

Sinowski, W. and Auerswald, K. (1999), "Using relief parameters in a discriminant analysis to stratify geological areas with different spatial variability of soil properties", *Geoderma,* vol. 89, no. 1-2, pp. 113-128.

Slaymaker, O. (2001), "The role of remote sensing in geomorphology and terrain analysis in the Canadian Cordillera", *International Journal of Applied Earth Observation and Geoinformation,* vol. 3, no. 1, pp. 11-17.

Smith, C. S., Howes, A. L., Price, B. and McAlpine, C. A. (2007), "Using a Bayesian belief network to predict suitable habitat of an endangered mammal – The Julia Creek dunnart (Sminthopsis douglasi)", *Biological Conservation,* vol. 139, no. 3-4, pp. 333-347.

Sommer, M., Wehrhan, M., Zipprich, M., Weller, U., zuCastell, W., Ehrich, S., Tandler, B. and Selige T. (2003), "Hierarchical data fusion for mapping soil units at field scale", *Geoderma,* vol. 11, pp. 179-196.

Stockwell, D. R. B. (2006), "Improving ecological niche models by data mining large environmental datasets for surrogate models", *Ecological Modelling,* vol. 192, no. 1-2, pp. 188-196.

Stoorvogel, J. J., Kempen, B., Heuvelink, G. B. M. and de Bruin, S. (2009), "Implementation and evaluation of existing knowledge for digital soil mapping in Senegal", *Geoderma,* vol. 149, no. 1-2, pp. 161-170.

Sturdy, R. G. (1971), *Soils in Essex 1 (Sheet TQ59 - Harold Hill),* , Harpenden.

Svendsen, J. I., Alexanderson, H., Astakhov, V. I., Demidov, I., Dowdeswell, J. A., Funder, S., Gataullin, V., Henriksen, M., Hjort, C., Houmark-Nielsen, M., Hubberten, H. W., Ingólfsson, Ó., Jakobsson, M., Kjær, K. H., Larsen, E., Lokrantz, H., Lunkka, J. P., Lyså, A., Mangerud, J., Matiouchkov, A., Murray, A., Möller, P., Niessen, F., Nikolskaya, O., Polyak, L., Saarnisto, M., Siegert, C., Siegert, M. J., Spielhagen, R. F. and Stein, R. (2004), "Late Quaternary ice sheet history of northern Eurasia", *Quaternary Science Reviews,* vol. 23, no. 11-13, pp. 1229-1271.

Thomas, A. L., Dambrine, E., King, D., Party, J. P. and Probst, A. (1999a), "A spatial study of the relationships between streamwater acidity and geology, soils and relief (Vosges, northeastern France)", *Journal of Hydrology,* vol. 217, no. 1-2, pp. 35-45.

Thomas, A. L., King, D., Dambrine, E., Couturier, A. and Roque, J. (1999b), "Predicting soil classes with parameters derived from relief and geologic materials in a sandstone region of the Vosges mountains (Northeastern France)", *Geoderma,* vol. 90, no. 3-4, pp. 291-305.

Toomanian, N., Jalalian, A., Khademi, H., Eghbal, M. K. and Papritz, A. (2006), "Pedodiversity and pedogenesis in Zayandeh-rud Valley, Central Iran", *Geomorphology,* vol. 81, no. 3-4, pp. 376-393.

Turenne, J. (2009), *Email Communication*, USDA-Natural Resources Conservation Service, Warwick, Rhode Island.

Tzortzis, M., Tsertos, H., Christofides, S. and Christodoulides, G. (2003), "Gamma-ray measurements of naturally occuring radioactive samples from Cyprus characteristic geological rocks", *Radiat.Measur.,* vol. 37, pp. 221-229.

USDA (2002), *Barnstable County, Massachusetts Soil Parent Materials (Surficial Geology)*, United States Department of Agriculture, Natural Resource Conservation Service, West Wareham, Massachusetts.

USDA Soil Conservation Service (1975), *General soil map of Hamilton County, New York*, Prepared for the Adirondack Park Agency by the U.S. Department of Agriculture, Soil Conservation Service in cooperation with Cornell University Agricultural Experiment Station., USA.

Viscarra Rossel, R. A., Cattle, S. R., Ortega, A. and Fouad, Y. (2009), "In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy", *Geoderma,* vol. 150, no. 3-4, pp. 253-266.

Visser, H. and de Nijs, T. (2006), "The Map Comparison Kit", *Environmental Modelling & Software,* vol. 21, no. 3, pp. 346-358.

West, I. (2007), *The Purbeck Formation (Upper Jurassic - Lower Cretaceous) of Southern England; a Geological Bibliography: Geology of the Wessex Coast.*

Internet field guide., available at: http://www.soton.ac.uk/~imw/Purbeck-Bibliography.htm. (accessed May 2007).

Wielemaker, W. G., de Bruin, S., Epema, G. F. and Veldkamp, A. (2001), "Significance and application of the multi-hierarchical landsystem in soil mapping", *Catena,* vol. 43, no. 1, pp. 15-34.

Wilford, J. R. (1992), "Regolith mapping using integrated Landsat TM imagery and high resolution gamma-ray spectrometric imagery - Cape York Peninsula", *Bureau of natural resources, Australia,* vol. 1992/78.

Wilford, J. R., Beirwirth, P. N. and Craig, M. A. (1997), "Application of airborne gamma-ray spectrometry in soil/regolith mapping and applied geomorphology", *AGSO Journal of Australian Geology and Geophysics,* vol. 17(2), pp. 201-216.

Won-In, K. and Charusiri, P. (2001), "Enhancement of thematic papper satellite images for geological mappin of ChoDien area, Northern Vietnam", *International Journal of Applied Earth Observation and Geoinformation,* vol. 4, pp. 183-193.

Worrall, L., Munday, T. J. and Green, A. A. (1999), "Airborne electromagnetics — Providing new perspectives on geomorphic process and landscape development in regolith-dominated terrains", *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy,* vol. 24, no. 10, pp. 855-860.

Wysocki, D. A., Schoeneberger, P. J. and LaGarry, H. E. (2005), "Soil surveys: a window to the subsurface", *Geoderma,* vol. 126, no. 1-2, pp. 167-180.

Yamada, K., Elith, J., McCarthy, M. and Zerger, A. (2003), "Eliciting and integrating expert knowledge for wildlife habitat modelling", *Ecological Modelling,* vol. 165, no. 2-3, pp. 251-264.

Zhang, X., Pazner, M. and Duke, N. (2007), "Lithologic and mineral information extraction for gold exploration using ASTER data in the south Chocolate Mountains (California)", *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 62, no. 4, pp. 271-282.

Zhu, A. (1997), "A similarity model for representing soil spatial information: Fuzzy Sets in Soil Science", *Geoderma,* vol. 77, no. 2-4, pp. 217-242.

Zhu, A., Band, L. E., Dutton, B. and Nimlos, T. J. (1996), "Automated soil inference under fuzzy logic: Fuzzy Modelling in Ecology", *Ecol.Model.,* vol. 90, no. 2, pp. 123-145.

APPENDICES

Appendix 1 – Study Area Maps

Appendix 2 – NSRI Soil Parent Material Classification

Appendix 3 – ESB Soil Parent Material Classification

Appendix 4 – Mapped Results

Appendix 5 - Corrections to Expector Equations

Appendix 6 - Parent materials and the predictive ability of evidence layers

Appendix 7 – Data processing for the derivation of the SLOPE layers

Appendix 8 – Soil Parent Material Profile Diagrams

Digital Appendix 1 – The Expert Knowledge Workbook

Digital Appendix 2 – The probability model workbook
(warning, running the model takes a number of hours).